



Goggle : People Video Analytics and Deep Learning Platform

Goggle แพลตฟอร์มการเรียนรู้เชิงลึกและระบบวิเคราะห์การกระทำของมนุษย์

นายปฐมพงศ์ สินธุจาม

นายศุภกร เบญจวิกรัย

นายอุกฤษฎ์ เลิศวรรณาการ

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตร

ปริญญาวิศวกรรมศาสตรบัณฑิต สาขาวิชาวิศวกรรมหุ่นยนต์และระบบอัตโนมัติ

สถาบันวิทยาการหุ่นยนต์ภาคสนาม

มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าธนบุรี

ปีการศึกษา 2562



Goggle : People Video Analytics and Deep Learning Platform

Goggle แพลตฟอร์มการเรียนรู้เชิงลึกและระบบวิเคราะห์การกระทำของมนุษย์

นายปฐมพงศ์ สินธุจาม

นายศุภกร เบญจวิกรัย

นายอุกฤษฎ์ เลิศวรรณาการ

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตร

ปริญญาวิศวกรรมศาสตรบัณฑิต สาขาวิชาชีวกรรมทุนยนต์และระบบอัตโนมัติ

สถาบันวิทยาการทุนยนต์ภาครสนา

มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าธนบุรี

ปีการศึกษา 2562

Google แพลตฟอร์มการเรียนรู้เชิงลึกและระบบวิเคราะห์การกระทำของมนุษย์

นายปฐมพงศ์ สินธุ์งาม

นายศุภกร เบญจวิกรัย

นายอุกฤษฎ์ เลิศวรณาการ

วิทยานิพนธ์เป็นส่วนหนึ่งของการศึกษาตามหลักสูตร

ปริญญาวิศวกรรมศาสตรบัณฑิต สาขาวิชาชีวกรรมหุ่นยนต์และระบบอัตโนมัติ

สถาบันวิทยาการหุ่นยนต์ภาคสนาม

มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าธนบุรี

ปีการศึกษา 2562

คณะกรรมการสอบวิทยานิพนธ์

ประธานกรรมการสอบวิทยานิพนธ์

(ดร.วรารสิณี ฉายแสงมงคล)

อาจารย์ที่ปรึกษาวิทยานิพนธ์

(ดร.วรารสิณี ฉายแสงมงคล)

อาจารย์ที่ปรึกษาวิทยานิพนธ์

()

กรรมการสอบวิทยานิพนธ์

(อ.บวรศักดิ์ ศกุลเกื้อภูลสุข)

กรรมการสอบวิทยานิพนธ์

(ดร.บุญทริกา เกษมสันติธรรม)

ลิขสิทธิ์ของมหาวิทยาลัยเทคโนโลยีพระจอมเกล้าธนบุรี

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

บทคัดย่อ

งานวิทยานิพนธ์นี้เป็นงานที่เกี่ยวกับการออกแบบและสร้างเครื่องมือกำกับคุณลักษณะด้วยปัญญาประดิษฐ์ รวมถึงระบบบุคลากรกระทำของมนุษย์ โดยใช้ชื่อว่า Goggle แพลตฟอร์มการเรียนรู้เชิงลึกและระบบบุคลากรกระทำของมนุษย์ ซึ่งมีจุดประสงค์เพื่อให้ผู้พัฒนาสามารถใช้งานเครื่องมือกำกับคุณลักษณะด้วยปัญญาประดิษฐ์ในการสร้างชุดข้อมูลสำหรับสร้างปัญญาประดิษฐ์ได้ง่ายและสะดวกขึ้น ภาพรวมของวิทยานิพนธ์นี้จะแบ่งออกเป็นหัวข้อหลักสองส่วน โดยที่ส่วนแรกเป็นการออกแบบและสร้างแอพพลิเคชันที่ใช้ในการสร้างชุดข้อมูลจากวิดีโอ สำหรับพัฒนาโมเดลปัญญาประดิษฐ์ และส่วนที่สองเป็นการออกแบบและสร้างระบบบุคลากรกระทำของมนุษย์ภายใต้ในสำนักงาน

คำสำคัญ : ระบบบุคลากรกระทำของมนุษย์ / เครื่องมือกำกับคุณลักษณะด้วยปัญญาประดิษฐ์ / Goggle

กิตติกรรมประกาศ

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

ขอขอบพระคุณอาจารย์อาจารย์ บวรศักดิ์ สกุลเกื้อกูลสุข ที่กรุณาให้เกียรติเป็นกรรมการสอบวิทยานิพนธ์ ให้คำแนะนำที่เป็นประโยชน์ต่อการวิจัย และการแก้ไขปรับปรุงงานวิจัย ตลอดจนตรวจสอบแก้วิทยานิพนธ์ให้ดำเนินไปอย่างสมบูรณ์

ขอขอบพระคุณอาจารย์ ดร.บุญทริกา เกษมสันติธรรม ที่กรุณาให้เกียรติเป็นกรรมการสอบวิทยานิพนธ์ ให้คำแนะนำที่เป็นประโยชน์ต่อการวิจัย และการแก้ไขปรับปรุงงานวิจัย ตลอดจนตรวจสอบแก้วิทยานิพนธ์ให้ดำเนินไปอย่างสมบูรณ์

ขอขอบพระคุณคณาจารย์ 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 | ๔ |
| สารบัญ | ๕ |
| รายการรูปภาพ | ๗ |
| รายการตาราง..... | ๘ |
| รายการสัญลักษณ์..... | ๙ |
| ประมวลศัพท์และตัวย่อ..... | ๙ |
| บทที่ 1 บทนำ..... | ๑ |
| 1.1 ที่มาและความสำคัญ..... | 1 |
| 1.2 วัตถุประสงค์..... | 1 |
| 1.3 ประโยชน์ที่คาดว่าจะได้รับ | 1 |
| 1.4 ขอบเขตการดำเนินงาน..... | 2 |
| 1.5 ขั้นตอนการดำเนินงาน | 2 |
| บทที่ 2 ทฤษฎี/การวิจัยที่เกี่ยวข้อง | 4 |
| 2.1 การวิเคราะห์ผลวิดีโอ..... | 4 |
| 2.1.1 การตรวจจับวัตถุ..... | 4 |
| 2.1.2 ระบบท่านายตำแหน่งถัดไปของวัตถุ..... | 5 |
| 2.1.3 ระบบระบุตัวตนของบุคคล | 6 |
| 2.1.4 ระบบจำแนกการกระทำ | 6 |
| 2.2 เครื่องมือสำหรับการวิเคราะห์ผลวิดีโอ | 14 |
| 2.2.1 โมเดลปัญญาประดิษฐ์สำหรับจำแนกการกระทำมุขย์..... | 14 |
| 2.2.2 เครื่องมือสำหรับสร้างชุดข้อมูล | 18 |
| 2.3 ทฤษฎีที่เกี่ยวข้อง | 20 |
| 2.3.1 Optical flow..... | 20 |
| บทที่ 3 ระเบียบวิธีวิจัย | 22 |
| 3.1 ความต้องการของระบบ..... | 22 |
| 3.1.1 ความต้องการเชิงการใช้งาน (functional requirements)..... | 22 |

สารบัญ (ต่อ)

| เรื่อง | หน้า |
|---|------|
| 3.1.2 ความต้องการเชิงวิศวกรรม (non-functional requirements)..... | 22 |
| 3.2 หน้าที่ความรับผิดชอบ..... | 23 |
| 3.3 เครื่องมือที่ใช้ในงานวิจัย | 23 |
| 3.4 ภาษาที่ใช้ในการพัฒนาระบบ | 24 |
| 3.5 Program library ที่ใช้ในการพัฒนาระบบและแอพพลิเคชัน | 24 |
| 3.6 แผนการดำเนินงาน | 24 |
| 3.7 ภาพรวมระบบของเครื่องมือสำหรับกำกับข้อมูลด้วยปัญญาประดิษฐ์..... | 25 |
| 3.8 การออกแบบหน้าต่างแอพพลิเคชันของเครื่องมือสำหรับกำกับข้อมูลด้วยปัญญาประดิษฐ์..... | 26 |
| 3.8.1 เครื่องมือสำหรับกำกับข้อมูลด้วยปัญญาประดิษฐ์..... | 26 |
| 3.9 การออกแบบการทดสอบการตรวจสอบจับวัดๆ..... | 36 |
| 3.9.1 ทดสอบประสิทธิภาพการทำงานของโมเดลปัญญาประดิษฐ์สำหรับการทำการตรวจจับภาพบุคคล | 36 |
| 3.10 การออกแบบการทดสอบการทำงานของระบบทำงานตามแบบต่อไปของมนุษย์..... | 37 |
| 3.10.1 ทดสอบประสิทธิภาพการทำงานของระบบทำงานตามแบบต่อไปของวัตถุในวิดีโอ | 37 |
| 3.11 การออกแบบการทดสอบการระบุตัวตนของมนุษย์ | 38 |
| 3.11.1 ทดสอบประสิทธิภาพการทำงานของระบบระบุตัวตนของบุคคลภายในภาพ | 38 |
| 3.12 การออกแบบการทดสอบการจดจำการกระทำการของมนุษย์ | 39 |
| 3.12.1 ทดสอบประสิทธิภาพการทำงานของโมเดลปัญญาประดิษฐ์ที่เคยถูกเทรน์ผ่าน AVA โดยใช้ชุดข้อมูลของ AVA ใน การทดสอบและเทียบผลลัพธ์กับแหล่งอ้างอิง..... | 39 |
| 3.12.2 ทดสอบประสิทธิภาพการทำงานของโมเดลปัญญาประดิษฐ์ที่เคยสร้างด้วย AVA และใช้ชุดข้อมูลที่ผู้วิจัยสร้างขึ้นในการทดสอบและเทียบผลลัพธ์กับแหล่งอ้างอิง..... | 40 |
| 3.12.3 ทดสอบประสิทธิภาพการทำงานของโมเดลปัญญาประดิษฐ์ที่ถูกสร้างด้วยชุดข้อมูลที่ผู้วิจัยสร้างขึ้น และใช้ชุดข้อมูลที่ผู้วิจัยสร้างขึ้นในการทดสอบและเทียบผลลัพธ์กับแหล่งอ้างอิง..... | 41 |
| บทที่ 4 ผลการดำเนินงาน | 42 |
| 4.1 Labeling tool | 42 |
| 4.1.1 หน้าต่างแสดงผลของแอพพลิเคชัน | 42 |
| 4.1.2 ผลลัพธ์การทำงานในแต่ละหน้าต่างของแอพพลิเคชัน | 46 |

สารบัญ (ต่อ)

| เรื่อง | หน้า |
|---|------|
| 4.2 ผลการทดลองการตรวจจับวัตถุ | 49 |
| 4.2.1 ข้อมูลรายละเอียดประกอบการทดสอบ | 49 |
| 4.2.2 ทดสอบประสิทธิภาพการทำงานของโมเดลปัญญาประดิษฐ์สำหรับการทำการตรวจจับภาพบุคคล | 49 |
| 4.2.3 ทดสอบประสิทธิภาพการทำงานของโมเดลปัญญาประดิษฐ์สำหรับการทำการตรวจจับภาพบุคคล | 49 |
| 4.3 ผลการทดสอบการทำนายตำแหน่งต่อไปของมนุษย์ | 51 |
| 4.3.1 ข้อมูลรายละเอียดประกอบการทดสอบ | 51 |
| 4.3.2 ทดสอบประสิทธิภาพ และความเร็วในการประมวลผล | 51 |
| 4.4 ผลการทดสอบระบบระบุตัวตนของมนุษย์ | 52 |
| 4.4.1 ทดสอบประสิทธิภาพการทำงานของโมเดลปัญญาประดิษฐ์สำหรับการระบุตัวตนของบุคคล | 52 |
| 4.5 ผลการทดสอบการจัดการกระทำของมนุษย์ | 54 |
| 4.5.1 ทดสอบประสิทธิภาพการทำงานของโมเดลปัญญาประดิษฐ์ที่เคยถูกเทรนด์ผ่าน AVA เทียบผลลัพธ์กับแหล่งอ้างอิง ได้ผลการทดลองดังตารางต่อไปนี้ | 54 |
| 4.5.2 ผลการทดสอบประสิทธิภาพการทำงานของโมเดลปัญญาประดิษฐ์ที่เคยถูกเทรนด์ผ่าน AVA และ ใช้ชุดข้อมูลที่ผู้วิจัยสร้างขึ้น ในการทดสอบและเทียบผลลัพธ์กับแหล่งอ้างอิง ... | 55 |
| 4.5.3 ทดสอบประสิทธิภาพการทำงานของโมเดลปัญญาประดิษฐ์ที่เคยถูกเทรนด์ผ่านชุดข้อมูลสำหรับการเทรนด์ที่ผู้วิจัยสร้างขึ้น และ ใช้ชุดข้อมูลที่ผู้วิจัยสร้างขึ้น ในการทดสอบและเทียบผลลัพธ์การทดสอบก่อนหน้า | 55 |
| เอกสารอ้างอิง | 57 |
| ภาคผนวก ก ตัวอย่างชุดข้อมูลที่ผู้วิจัยสร้างขึ้น | 58 |

รายการรูปภาพ

| รูป | หน้า |
|--|------|
| รูปที่ 2.1 แนวคิดของระบบทำนายตำแหน่งตัวแทนงัดไปของวัตถุ | 5 |
| รูปที่ 2.2 หลักการของ Residual block ของ ResNet | 14 |
| รูปที่ 2.3 โครงสร้างของโมเดลปัญญาประดิษฐ์ I3D ^[4] | 16 |
| รูปที่ 2.4 โครงสร้างของโมเดลปัญญาประดิษฐ์ I3D ^[4] | 17 |
| รูปที่ 2.5 UI ของโปรแกรม DarkLabel | 18 |
| รูปที่ 2.6 UI ของโปรแกรม OpenLabeling | 19 |
| รูปที่ 2.7 ตัวอย่างการเคลื่อนที่ของกลุ่มบล็อก | 20 |
| รูปที่ 3.1 ภาพรวมระบบของเครื่องมือสำหรับกำกับข้อมูลด้วยปัญญาประดิษฐ์ | 25 |
| รูปที่ 3.2 กระบวนการหลักของเครื่องมือสำหรับกำกับข้อมูลด้วยปัญญาประดิษฐ์ | 26 |
| รูปที่ 3.3 หน้าต่าง Select ของเครื่องมือสำหรับกำกับข้อมูลด้วยปัญญาประดิษฐ์ | 27 |
| รูปที่ 3.4 ตำแหน่งของแต่ละวิดเจ็ตในหน้าต่าง Select | 28 |
| รูปที่ 3.5 หน้าต่าง Detect ของเครื่องมือสำหรับกำกับข้อมูลด้วยปัญญาประดิษฐ์ | 29 |
| รูปที่ 3.6 ตำแหน่งของแต่ละวิดเจ็ตในหน้าต่าง Detect | 30 |
| รูปที่ 3.7 หน้าต่าง Track ของเครื่องมือสำหรับกำกับข้อมูลด้วยปัญญาประดิษฐ์ | 31 |
| รูปที่ 3.8 ตำแหน่งของแต่ละวิดเจ็ตในหน้าต่าง Track | 32 |
| รูปที่ 3.9 หน้าต่าง Label ของเครื่องมือสำหรับกำกับข้อมูลด้วยปัญญาประดิษฐ์ | 33 |
| รูปที่ 3.10 ตำแหน่งของแต่ละวิดเจ็ตในหน้าต่าง Label | 34 |
| รูปที่ 3.11 ตัวอย่างข้อมูลภายในไฟล์ xml | 35 |
| รูปที่ 4.1 รูปหน้าต่างแสดงผลของหน้าต่าง Select | 42 |
| รูปที่ 4.2 รูปหน้าต่างแสดงผลของหน้าต่าง Detect | 43 |
| รูปที่ 4.3 รูปหน้าต่างแสดงผลของหน้าต่าง Track | 44 |
| รูปที่ 4.4 รูปหน้าต่างแสดงผลของหน้าต่าง Label | 45 |
| รูปที่ 4.5 รูปผลลัพธ์การแยกเฟรมที่มีมนุษย์อยู่ และไม่มีมนุษย์อยู่ภายในเฟรม | 46 |
| รูปที่ 4.6 รูปคีย์เฟรมที่ถูกตีกรอบสีเหลืองในส่วนที่มีมนุษย์อยู่ | 46 |
| รูปที่ 4.7 รูปผลลัพธ์การทำงานของหน้าต่าง Track | 47 |
| รูปที่ 4.8 รูปผลลัพธ์การทำงานของหน้าต่าง Label | 48 |
| รูปที่ 4.9 ภาพตัวอย่างชุดข้อมูลสำหรับการทำงานทดลองครั้งที่ 1 | 52 |

รายการรูปภาพ (ต่อ)

| รูป | หน้า |
|---|------|
| รูปที่ 4.10 ภาพตัวอย่างชุดข้อมูลสำหรับการทดลองครั้งที่ 2..... | 52 |
| รูปที่ 4.11 ภาพตัวอย่างชุดข้อมูลสำหรับการทดลองครั้งที่ 3..... | 54 |
| รูปที่ ก.1 รูปผลลัพธ์การทำงานของหน้าต่าง Track..... | 58 |

รายการตาราง

| ตาราง | หน้า |
|---|------|
| ตารางที่ 1.1 แผนการดำเนินงาน | 3 |
| ตารางที่ 2.1 ผลการทดสอบโมเดลต่างๆบนชุดข้อมูลสำหรับทดสอบของ YouTube-8M..... | 8 |
| ตารางที่ 2.2 ผลการทดสอบโมเดลต่างๆบนชุดข้อมูลสำหรับทดสอบของ Sports-1M..... | 9 |
| ตารางที่ 2.3 ผลการทดสอบโมเดลต่างๆบนชุดข้อมูลสำหรับทดสอบของ ActivityNet..... | 9 |
| ตารางที่ 2.4 ผลการทดลองของวิธีต่างๆบนคุณลักษณะระดับเฟรม | 11 |
| ตารางที่ 2.5 ประสิทธิภาพของโมเดล Resnet50 I3D ที่ใช้ชุดข้อมูล Kinetics และ Moments in Time .. | 13 |
| ตารางที่ 2.6 อัตราเร้อยลของความผิดพลาดของชุดข้อมูลทดสอบ ImageNet..... | 14 |
| ตารางที่ 2.7 ค่าความผิดพลาดที่ได้จากการทดลองจำนวนขั้นของโมเดลปัญญาประดิษฐ์ ResNet บนชุดของข้อมูล CIFAR-10 | 15 |
| ตารางที่ 2.8 ประสิทธิภาพของโมเดล I3D แบบ two-stream เมื่อใช้ข้อมูลจาก UCF-101, HMDB-51 และ Kinetics ใน การสร้างและทดสอบด้วยเครื่องมือวัดผลแบบความแม่นยำจากการทำนายอันดับแรก | |
| ตารางที่ 4.1 ช้อมูลผลการทำงานของโมเดลปัญญาประดิษฐ์สำหรับการทำการตรวจจับภาพบุคคล | 17 |
| อ้างอิงข้อมูลจากเว็บไซต์ของ yolo | 49 |
| ตารางที่ 4.2 ช้อมูลผลการทำงานของโมเดลปัญญาประดิษฐ์สำหรับการทำการตรวจสอบภาพบุคคลหลังจากการทดลอง | 50 |
| ตารางที่ 4.3 ผลการทดสอบประสิทธิภาพของการตรวจจับกรอบสี่เหลี่ยมภายในวิดีโอ..... | 51 |
| ตารางที่ 4.4 ผลการทดสอบความแม่นยำของโมเดลปัญญาประดิษฐ์..... | 52 |
| ตารางที่ 4.5 ผลการทดสอบความแม่นยำสำหรับการระบุบุคคลของโมเดลปัญญาประดิษฐ์ครั้งที่ 1 | 52 |
| ตารางที่ 4.6 ผลการทดสอบความแม่นยำสำหรับการระบุบุคคลของโมเดลปัญญาประดิษฐ์ครั้งที่ 2 | 53 |
| ตารางที่ 4.7 ผลการทดสอบความแม่นยำสำหรับการระบุบุคคลของโมเดลปัญญาประดิษฐ์ครั้งที่ 3 | 54 |
| ตารางที่ 4.8 ผลการทดสอบความแม่นยำของโมเดลปัญญาประดิษฐ์เทียบผลลัพธ์กับแหล่งอ้างอิง..... | 54 |
| ตารางที่ 4.9 ผลการทดสอบความแม่นยำของโมเดลปัญญาประดิษฐ์ เมื่อใช้กับชุดข้อมูลที่ผู้วิจัยสร้างขึ้น.... | 55 |
| ตารางที่ 4.10 ผลการทดสอบความแม่นยำของโมเดลปัญญาประดิษฐ์ที่ผู้วิจัยสร้างขึ้น ใช้กับชุดข้อมูลที่ผู้วิจัยสร้างขึ้น | 55 |

รายการสัญลักษณ์

| | |
|----------|--------------|
| θ | เชิงตัว |
| d | distance |
| kg | Kilogram |
| m^2 | Square Metre |

ประมวลศัพท์และตัวย่อ

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

บทที่ 1

บทนำ

1.1 ที่มาและความสำคัญ

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

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

1.2 วัตถุประสงค์

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

1.3 ประโยชน์ที่คาดว่าจะได้รับ

- เพิ่มความสะดวกในการสร้างชุดข้อมูลสำหรับพัฒนาโมเดลปัญญาประดิษฐ์จากวิดีโอ
- ต้นแบบระบบบวิเคราะห์วิดีโอที่สามารถจำแนกการกระทำการของมนุษย์ภายในสำนักงานได้

1.4 ขอบเขตการดำเนินงาน

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

1.5 ขั้นตอนการดำเนินงาน

การดำเนินงานวิจัยถูกแบ่งออกเป็นสามส่วน โดยส่วนแรกคือการศึกษาเทคโนโลยีในปัจจุบันเพื่อหาความเป็นไปได้และกำหนดขอบเขตของงาน ส่วนที่สองคือเครื่องมือกำกับคุณลักษณะด้วยปัญญาประดิษฐ์ เป็นส่วนที่ออกแบบและสร้างเครื่องมือสำหรับช่วยผู้พัฒนาในการสร้างชุดข้อมูล และส่วนที่สุดท้ายคือการนำชุดข้อมูลที่ได้จากการใช้เครื่องมือกำกับคุณลักษณะด้วยปัญญาประดิษฐ์ไปพัฒนาโมเดลปัญญาประดิษฐ์สำหรับการจำแนกการกระทำการของมนุษย์ภายในสำนักงาน

ศึกษาค้นคว้าเอกสารและงานวิจัยที่เกี่ยวข้อง

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

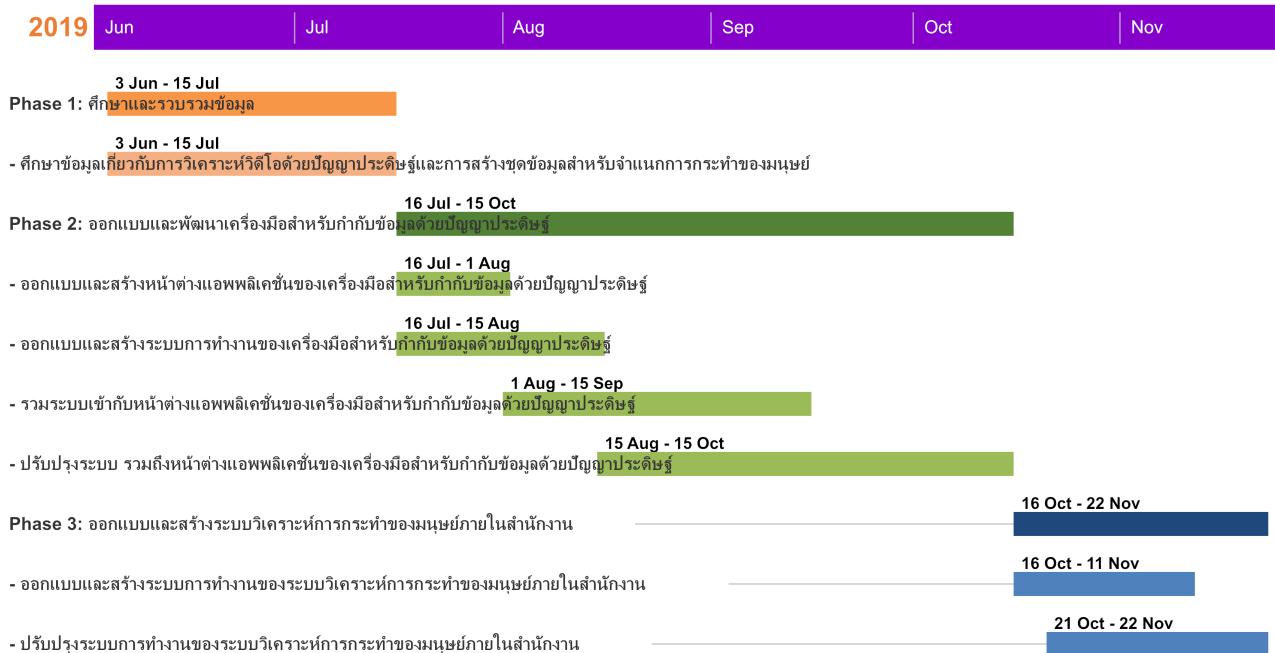
เครื่องมือกำกับคุณลักษณะด้วยปัญญาประดิษฐ์

1. ออกแบบและสร้างหน้าต่างแอพพลิเคชันของเครื่องมือกำกับคุณลักษณะด้วยปัญญาประดิษฐ์
2. ออกแบบและสร้างระบบของเครื่องมือกำกับคุณลักษณะด้วยปัญญาประดิษฐ์
3. ทดสอบและปรับปรุงการทำงานของเครื่องมือกำกับคุณลักษณะด้วยปัญญาประดิษฐ์

โมเดลปัญญาประดิษฐ์สำหรับการจำแนกการกระทำการของมนุษย์ภายในสำนักงาน

1. สร้างชุดข้อมูลสำหรับสร้างโมเดลปัญญาประดิษฐ์จากเครื่องมือกำกับคุณลักษณะด้วยปัญญาประดิษฐ์
2. สร้างโมเดลปัญญาประดิษฐ์สำหรับการจำแนกการกระทำการของมนุษย์ภายในสำนักงาน
3. ทดสอบโมเดลปัญญาประดิษฐ์สำหรับการจำแนกการกระทำการของมนุษย์ภายในสำนักงาน

แผนการดำเนินงาน



ตารางที่ 1.1: แผนการดำเนินงาน

บทที่ 2

ทฤษฎี/การวิจัยที่เกี่ยวข้อง

การวิเคราะห์วิดีโoinปัจจุบันนั้นมีวิธีและเทคนิคหลากหลาย ผู้วิจัยจึงต้องศึกษาองค์ความรู้และงานวิจัยที่เกี่ยวข้องกับวัตถุประสงค์ของงาน เพื่อศึกษาและใช้เป็นแนวทางในการประยุกต์สำหรับสร้างเครื่องมือสำหรับกำกับข้อมูลด้วยปัญญาประดิษฐ์ และโมเดลปัญญาประดิษฐ์สำหรับการจำแนกการกระทำของมนุษย์ ซึ่งหัวข้อที่ผู้วิจัยได้ไปศึกษามา มีดังต่อไปนี้

1. การวิเคราะห์ผลวิดีโอ

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

2. เครื่องมือสำหรับการวิเคราะห์ผลวิดีโอ

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

3. ทฤษฎีที่เกี่ยวข้อง

- (a) Optical flow

2.1 การวิเคราะห์ผลวิดีโอ

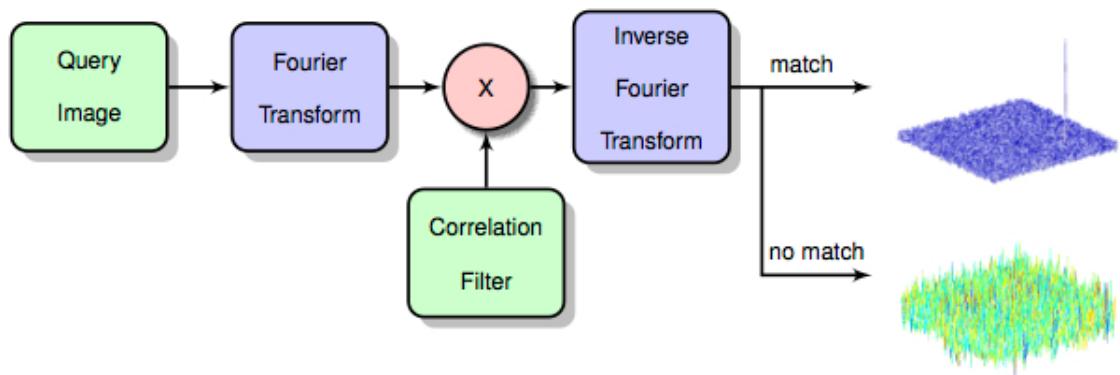
ในส่วนของงานวิจัยสิ่งที่สนใจ คือ ข้อมูลการกระทำของมนุษย์แต่ละคนภายในวิดีโอ เพื่อที่จะได้ผลลัพธ์ที่มีประสิทธิภาพอย่างมากเป็นข้อมูลของสิ่งที่สนใจ เช่น จำนวนคนที่เดินผ่านกล้อง หรือทิศทางการเดินของคนในวิดีโอ จึงจำเป็นต้องใช้การวิเคราะห์ผลวิดีโอเพื่อที่จะสกัดสิ่งที่สนใจออกมาจากวิดีโอ ซึ่งการวิเคราะห์ผลวิดีโอมีหลากหลายกระบวนการ โดยในแต่ละกระบวนการจะมีจุดประสงค์ของการทำและผลลัพธ์หลังการประมวลผลที่แตกต่างกัน ในหัวข้อนี้จะมาอธิบายถึงกระบวนการในการวิเคราะห์ผลของวิดีโอและผลลัพธ์ของกระบวนการนั้น

2.1.1 การตรวจจับวัตถุ

การตรวจจับวัตถุนั้นเป็นกระบวนการที่นิยมใช้ในการวิเคราะห์ผลของวิดีโอ กล่าวคือกระบวนการที่ผู้วิจัยจะต้องทำคือระบุสิ่งที่สนใจว่าอยู่ตำแหน่งใด ซึ่งในปัจจุบันการทำการตรวจจับวัตถุมักนำปัญญาประดิษฐ์มาใช้เนื่องจากมีประสิทธิภาพสูง ซึ่งโมเดลปัญญาประดิษฐ์ที่เลือกใช้คือ YOLO-v3 320 เนื่องจากมีความสามารถตรวจจับวัตถุได้อย่างรวดเร็วและแม่นยำ โดยหลักการทำงานของ YOLO-v3 320 คือ นำรูปภาพที่ต้องการผ่านโครงข่ายประสาทเทียม (neural network) จากนั้นโครงข่ายจะแบ่งรูปภาพเป็นพื้นที่หลายๆ ส่วนแล้วนำความน่าจะเป็นของวัตถุในพื้นที่ว่าเป็นวัตถุใด สุดท้ายจะเลือกรอบสี่เหลี่ยมที่มีค่าคะแนนความน่าจะเป็นมากที่สุด

2.1.2 ระบบที่นำมายำต์แหน่งถัดไปของวัตถุ

การนำมายำต์แหน่งถัดไปของวัตถุ^[5] คือระบบที่ใช้สำหรับการติดตามการเคลื่อนไหวของวัตถุที่สนใจที่อยู่ในรูปภาพ โดยใช้การคำนวณทางคณิตศาสตร์ และการประมวลผลภาพ (image processing) ทำให้การประมวลผลนั้นเร็วกว่าการใช้โมเดลปัญญาประดิษฐ์



รูปที่ 2.1: แนวคิดของระบบที่นำมายำต์แหน่งถัดไปของวัตถุ

จากรูปที่ 2.1 เป็นหลักการในการนำมายำต์แหน่งต่อไป โดยการนำรูปมาผ่านกระบวนการแปลงฟูรีเยร์ (fourier transform) และนำมาร่วมกับ correlation filter ซึ่งเป็นตัวรองที่ใช้สำหรับการหาความสัมพันธ์กับวัตถุในภาพ จากนั้นทำการแปลงฟูรีเยร์กลับ (inverse fourier transform) เพื่อตรวจสอบว่าวัตถุในภาพนั้นอยู่ที่ตำแหน่งใด โดยมีการคำนวณเริ่มจากการหา correlation filter ที่ดีที่สุดโดยใช้วิธีลดผลรวมของข้อผิดพลาดกำลังสองให้น้อยที่สุดดังนี้

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

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 = อัตราการเรียนรู้

\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} = การแปลงฟรีเยอร์ผกผันแบบ逆 (inverse discrete fourier transform)

Z = พื้นที่สี่เหลี่ยมของวัตถุที่สนใจที่ได้จากการหาผังคุณลักษณะของภาพใหม่

โดยค่าของ y ที่ได้ออกมาจะทำให้รู้ถึงตำแหน่งของวัตถุที่สนใจได้ ณ ตำแหน่งที่ y มีค่าสูงสุด

2.1.3 ระบบระบุตัวตนของบุคคล

ระบบระบุตัวตนของบุคคล คือการระบุตัวตนของบุคคลภายในวิดีโอหรือระหว่างรูป 2 รูป สามารถนำมาประยุกต์ใช้ในด้านของการรักษาความปลอดภัย การตามหาบุคคล หรือการตรวจสอบการกระทำของบุคคลนั้นในวิดีโอด้วย การระบุตัวตนของบุคคลนั้นเป็นปัญหาที่ท้าทาย เนื่องจากคุณลักษณะทั่วไปของบุคคลในรูปไม่เพียงพอต่อการระบุบุคคลภายในรูปว่าเป็นบุคคลคนเดียวกันได้ ซึ่งวิธีการที่ใช้สำหรับการระบุตัวตนของบุคคลเรียกว่า Dynamically Matching Local Information (DMLI) ที่สามารถจัดแนวลายละเอียดข้อมูลของรูป และเพิ่มประสิทธิรูปให้สูงขึ้น

การทำงานของระบบระบุตัวตนของบุคคลจะเริ่มจากการแบ่งรูปออกเป็น 8 ส่วนและนำคุณลักษณะของรูปมาผ่านกระบวนการ normalization เพื่อลดความซ้ำซ้อนของข้อมูล แล้วนำมาเปรียบเทียบความแตกต่างของคุณลักษณะของรูป หลังจากนั้นหากค่าเฉลี่ยของความแตกต่างออกมา ถ้าค่าที่ออกมากใกล้เคียงกับ 0 จะหมายถึงบุคคลในรูปทั้งสองเป็นบุคคลเดียวกัน

2.1.4 ระบบจำแนกการกระทำ

ระบบจำแนกการกระทำเป็นกระบวนการสำหรับนำเสนาระบบทามนิยมหรือสิ่งที่สนใจอื่นๆที่เกิดการกระทำขึ้นภายในวิดีโอ โดยในหัวข้อนี้จะกล่าวถึงตั้งแต่ขั้นตอนการได้มาซึ่งชุดข้อมูลมีกระบวนการอย่างไร การนำโมเดลปัญญาประดิษฐ์มาใช้ในการจำแนกการกระทำ และการวัดผลของโมเดลปัญญาประดิษฐ์ โดยชุดข้อมูลที่ผู้วิจัยได้เลือกนำมาศึกษาจากชุดข้อมูลที่ถูกเป็นที่กล่าวถึงในปัจจุบัน และมีขนาดของชุดข้อมูลที่ใหญ่

จากข้อความข้างต้นชุดข้อมูลที่ผู้วิจัยได้เลือกนำมาใช้ได้แก่ YouTube-8M, AVA, Moment in Time โดยแต่ละชุดข้อมูลจะมีความแตกต่างกันในหลายๆด้าน แต่จะมีสิ่งที่เหมือนกัน คือ เป็นชุดข้อมูลสำหรับการวิเคราะห์วิดีโอที่สนใจการกระทำของมนุษย์ โดยในบทความนี้จะกล่าวถึงความแตกต่างในด้านต่างๆ เช่น เป้าหมายของแต่ละชุดข้อมูล วิธีการเก็บข้อมูลสำหรับชุดข้อมูล วิธีการสร้างคำกำกับคุณลักษณะ และรายละเอียดของชุดข้อมูล จากนั้นจะสรุปข้อมูลของแต่ละชุดข้อมูล

ชุดข้อมูล YouTube-8M

1. รายละเอียดของชุดข้อมูล

- (a) เป้าหมายของชุดข้อมูล : เพื่อจำแนกสาระสำคัญของวิดีโอ (video theme)
- (b) จำนวนของวิดีโอ : 8,264,650 วิดีโอ
- (c) ความยาวเฉลี่ยของแต่ละวิดีโอ : 229.6 วินาที
- (d) จำนวนของหมวดหมู่ของคำกำกับคุณลักษณะ : 4,800 หมวดหมู่
- (e) กฎในการรวบรวมวิดีโอดังนี้
 - i. ทุกๆหัวข้อต้องเป็นรูปธรรม
 - ii. ในแต่ละหัวข้อต้องมีจำนวนวิดีโอย่ำกกว่า 200 วิดีโอ
 - iii. ความยาวของวิดีโอดังนี้อยู่ระหว่าง 120 - 500 วินาที

หลังจากได้กฎในการรวบรวมวิดีโอลแล้ว ขั้นตอนต่อไปคือการสร้างคำศัพท์ที่ใช้ในการค้นหาข้อมูลวิดีโອจากใน YouTube

(f) ขั้นตอนในการสร้างคำศัพท์มีดังนี้

- i. กำหนดบัญชีขาว (whitelist) ของหัวข้อที่เป็นรูปธรรมมา 25 ชนิด เช่น กีฬา เป็นต้น
- ii. กำหนดบัญชีดำ (blacklist) ของหัวข้อที่คิดว่าไม่เป็นรูปธรรมไว้ เช่น software เป็นต้น
- iii. รวบรวมหัวข้อที่มีอยู่ในรายการที่อนุญาตอย่างน้อย 1 หัวข้อ และต้องไม่มีอยู่ในบัญชีดำซึ่งจะทำให้ได้หัวข้อที่ต้องการมาประมาณ 50,000 หัวข้อ
- iv. จากนั้นใช้ผู้ประเมินจำนวน 3 คน ในการคัดหัวข้อที่คิดว่าเป็นรูปธรรม และสามารถจดจำหรือเข้าใจได้เจ้ายโดยไม่ต้องเชี่ยวชาญในด้านนั้นๆ ซึ่งผู้ประเมิน ก็จะมีคำถามว่า “มันยากขนาดไหนถึงจะระบุได้ว่ามีหัวข้อดังกล่าวอยู่ในรูปหรือวิดีโอ โดยใช้เพียงแค่การมองเท่านั้น?” โดยแบ่งเป็นระดับดังนี้

- A. บุคคลทั่วไปสามารถเข้าใจได้
- B. บุคคลทั่วไปที่ผ่านการอ่านบทความที่เกี่ยวข้องมาแล้วสามารถเข้าใจได้
- C. ต้องเชี่ยวญในด้านใดซักด้านจึงจะเข้าใจได้
- D. เป็นไปไม่ได้ ถ้าไม่มีความรู้ที่ไม่ได้เป็นรูปธรรม
- E. ไม่เป็นรูปธรรม

- v. หลังจากคำถามข้างบนและการให้คะแนน จะทำการเก็บไว้เฉพาะหัวข้อที่มีคะแนนเฉลี่ยมากที่สุดอยู่ที่ประมาณ 2.5 คะแนนหรือต่ำกว่าเท่านั้น

- vi. ทำให้สุดท้ายเหลือเพียงประมาณ 10,000 หัวข้อที่สามารถใช้ได้

- vii. หลังจากได้หัวข้อที่คิดว่าเป็นรูปธรรมแล้วก็นำไปค้นหาและรวบรวมด้วย YouTube annotation system โดยมีขั้นตอนดังนี้

- A. สุ่มเลือกวิดีโอมาก 10 ล้านวิดีโอ พร้อมกับหัวข้อของวิดีโอ โดยใช้กฎที่กำหนดไว้ เอาหัวข้อที่มีจำนวนวิดีโอย่ำกกว่า 200 วิดีโอออก
- B. ทำให้เหลือจำนวนวิดีโอยู่ 8,264,650 วิดีโอ
- C. แยกออกเป็น 3 ส่วนคือ ชุดข้อมูลสำหรับสร้างโมเดล (train set) ชุดข้อมูลสำหรับตรวจสอบ (validate set) และชุดข้อมูลสำหรับทดสอบ (test set) ในอัตราส่วน 70:20:10 ตามลำดับ

2. โมเดลปัญญาประดิษฐ์

(a) การเตรียมข้อมูล

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

(b) โมเดลปัญญาประดิษฐ์ที่ใช้ในการทดสอบชุดข้อมูลแบบที่เป็นคุณลักษณะระดับเฟรม

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

(c) โมเดลปัญญาประดิษฐ์ที่ใช้ในการทดสอบชุดข้อมูลแบบที่เป็นคุณลักษณะระดับวิดีโอ

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

(d) เครื่องมือที่ใช้วัดผลสำหรับงานวิจัยนี้ คือ

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

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

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

ตารางที่ 2.1: ผลการทดสอบโมเดลต่างๆบนชุดข้อมูลสำหรับทดสอบของ YouTube-8M

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

| Approach | mAP | Hit@1 | PERR |
|---|------|-------------|-------------|
| Logistic regression | 58.0 | 60.1 | 79.6 |
| Mixture-of-2-experts | 61.3 | 63.2 | 82.6 |
| LSTM | 67.6 | 65.7 | 86.2 |
| Hierarchical 3D convolutions ^[8] | - | 61.0 | 80.0 |
| Stacked 3D convolutions ^[14] | - | 61.0 | 85.0 |
| LSTM with optical flow and pixels ^[13] | - | 73.0 | 91.0 |

ตารางที่ 2.2: ผลการทดสอบโมเดลต่างๆบนชุดข้อมูลสำหรับทดสอบของ Sports-1M

(g) ประสิทธิภาพของโมเดลปัญญาประดิษฐ์ที่ใช้ชุดข้อมูลของ YouTube-8M ในการสร้างเทียบกับชุดข้อมูลสำหรับทดสอบของ ActivityNet

| Approach | mAP | Hit@1 | Hit@5 |
|-----------------------------------|-------------|-------------|-------------|
| Mixture-of-2-experts | 77.6 | 74.9 | 91.6 |
| LSTM | 57.9 | 63.4 | 81.0 |
| Ma, Bargal et al. ^[10] | 53.8 | - | - |
| Heilbron et al. ^[3] | 43.0 | - | - |

ตารางที่ 2.3: ผลการทดสอบโมเดลต่างๆบนชุดข้อมูลสำหรับทดสอบของ ActivityNet

(h) ปัญหาที่พบ

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

ชุดข้อมูล Atomic visual action (AVA)

1. รายละเอียดชุดข้อมูล

- (a) เป้าหมายของชุดข้อมูล : สนใจการกระทำของมนุษย์เป็นศูนย์กลาง
 - (b) จำนวนของวิดีโอ : 640 วิดีโอ
 - (c) ความยาวเฉลี่ยของแต่ละวิดีโอ : 15 นาที และถูกสุมตัวอย่างด้วยความถี่ 1 hz หรือ 1 เฟรมต่อ วินาที
 - (d) จำนวนของหมวดหมู่ : 80 หมวดหมู่
 - (e) ขั้นตอนการเก็บข้อมูลสำหรับการทำชุดข้อมูลมีขั้นตอนการทำ 5 ขั้นดังนี้
 - i. การสร้างคำศัพท์การกระทำจะมีหลักการ 3 ข้อในการรวบรวมคำศัพท์ดังนี้
 - A. เก็บรวบรวมคำศัพท์ทั่วไปที่เกิดขึ้นในชีวิตประจำวัน
 - B. จะต้องมีเอกสารกักษณ์สามารถเห็นได้ชัดเจน เช่น การถือของ
 - C. กำหนดรูปแบบของคำศัพท์ขึ้นมา และใช้ความรู้จากชุดข้อมูลอื่นในการทำให้ได้หมวดหมู่การกระทำของมนุษย์ที่ครอบคลุม
 - ii. ภาพยนต์และส่วนที่เลือกมาใช้ทำชุดข้อมูล AVA ทั้งหมดจะถูกนำมาจาก YouTube โดยเริ่มจากการรวบรวมรายการซื้อของนักแสดงที่มีชื่อเสียง ซึ่งจะมีความหลากหลายของเชื้อชาติ รวมกันอยู่ วิดีโอที่ถูกคัดเลือกจะมีเกณฑ์ดังนี้
 - A. วิดีออด้วยในหมวด ภาพยนต์ และละครโทรทัศน์
 - B. วิดีโอจะต้องมีความยาวมากกว่า 30 นาที
 - C. เผยแพร่มาแล้วเป็นระยะเวลาน้อยกว่า 1 ปี
 - D. มีจำนวนยอดคนดูมากกว่า 1,000 ครั้ง
 - E. ลงทะเบียนวิดีโอบางประเภท เป็นภาพขาว-ดำ มีความละเอียดต่ำ การ์ตูน หรือวิดีโอกวน
 - iii. การสร้างกรอบสีเหลี่ยมครอบมนุษย์ที่อยู่ภายใต้ภาพประกอบด้วย 2 ขั้นตอน
 - A. สร้างกรอบสีเหลี่ยมโดยใช้โมเดลปัญญาประดิษฐ์ faster RCNN สำหรับการตรวจจับมนุษย์
 - B. ใช้มนุษย์ในการตรวจสอบและแก้ไขกรอบสีเหลี่ยมที่ผิดพลาด
 - iv. การเข้ามบุคคลในช่วงระยะเวลาสั้นๆของเฟรม ทำการเข้มกรอบสีเหลี่ยมที่อยู่ในช่วงเวลาเดียวกันด้วยวิธีการทำนายตำแหน่งโดยใช้模型มนุษย์ เป็นศูนย์กลาง ซึ่งจะนำมารាមความใกล้เคียงกันของคู่กรอบสีเหลี่ยม และใช้ person embedding จากนั้นจะใช้อัลกอริทึม Hungarian ในการหาตัวเลือกที่ดีที่สุด
 - v. การสร้างคำกำกับคุณลักษณะ
- การสร้างคำกำกับของการกระทำจะใช้มนุษย์ในการสร้าง ซึ่งจะใช้หน้าต่างโปรแกรมสำหรับช่วยเหลือในการสร้างซึ่งในหนึ่งกรอบสีเหลี่ยม สามารถมีคำกำกับของการกระทำได้สูงสุดถึง 7 คำ นอกจากนั้นสามารถตั้งสถานะเนื้อหาที่ไม่เหมาะสม หรือกรอบสีเหลี่ยมที่ผิดพลาดได้ อีกด้วย ในทางปฏิบัติจะสังเกตได้ว่ามันมีโอกาสผิดอย่างหลีกเลี่ยงไม่ได้ เมื่อต้องได้รับคำสั่งให้หาคำกำกับของการกระทำที่ถูกต้องจาก 80 หมวดหมู่ จึงแบ่งขั้นตอนออกเป็น 2 ขั้นตอน ดังนี้

- A. สร้างข้อเสนอสำหรับคำจำกัดของกระบวนการทำจากนั้นจับกลุ่มเข้าด้วยกัน ซึ่งจะทำให้มีโอกาสสูงต้องมากกว่าเป็นคำจำกัดแยกเดียว
 B. ผู้ตรวจสอบข้อเสนอจะตรวจสอบข้อเสนอที่ได้จากการชั้นตอนแรก ซึ่งในแต่ละวิดีโอดูให้มนุษย์ในการตรวจสอบ 3 คน เมื่อคำจำกัดของกระบวนการทำถูกตรวจสอบด้วยผู้ตรวจสอบข้อเสนออย่างน้อย 2 คน คำจำกัดของกระบวนการทำนั้นจะถูกยึดเป็นคำจำกัดหลัก

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 คือค่าที่ใช้วัดความสอดคล้องระหว่างสองเฟรมใช้สำหรับการวัดผลระดับเฟรม โดยจะเป็นหาพื้นที่ร่วมกันระหว่างกรอบสีเหลี่ยมที่ตรวจเจอและกรอบสีเหลี่ยมจริงของวัตถุ
 - ii. ค่า 3D IoUs คือค่าที่ใช้วัดความสอดคล้องระหว่างสองวิดีโอใช้สำหรับการวัดผลระดับวิดีโอด้วยเทียบกันระหว่างตำแหน่งคำตอบจริงในช่วงของเฟรมที่ต่อกัน (ground-truth tubes) และตำแหน่งคำตอบจากการตรวจจับในช่วงของเฟรมที่ต่อกัน (linked detection tubes) ซึ่งตำแหน่งคำตอบจริงในช่วงของเฟรมที่ต่อกัน หมายถึงการนำเอกสารอบสีเหลี่ยมจริงของวัตถุในเฟรมที่ติดกันมาเรียงต่อกัน และตำแหน่งคำตอบจากการตรวจจับในช่วงของเฟรมที่ต่อกัน หมายถึงการนำเอกสารอบสีเหลี่ยมที่ตรวจเจอนามาเรียงต่อกันในเฟรมที่ติดกันมาเรียงต่อกัน

(c) ประสิทธิภาพของโมเดลปัญญาประดิษฐ์ในปัจจุบัน

 - i. จากการทดสอบการเทียบโมเดลปัญญาประดิษฐ์ของงานวิจัยนี้และวิธีการอื่นๆ โดยนำไปทดสอบกับชุดข้อมูลวิดีโอดู 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: ผลการทดลองของวิธีต่างๆบนคุณลักษณะระดับเฟรม

- (d) ปัญหาที่พบ ในปัจจุบันยังไม่มีโมเดลปัญญาประดิษฐ์ที่ทดสอบด้วยชุดข้อมูล AVA และได้ผลการทำงานที่ดี เนื่องจากชุดข้อมูลนี้สนใจการกระทำของมนุษย์ที่มีรายละเอียดเล็กๆน้อยๆ ทำให้ยากต่อการทำนายสำหรับโมเดลปัญญาประดิษฐ์

ชุดข้อมูล Moments in Time

1. รายละเอียดชุดข้อมูล

- (a) เป้าหมายของชุดข้อมูล : สนับสนุนการที่เกิดขึ้นในวิดีโอ เช่น การกระทำของคนหรือสัตว์ เหตุการณ์ และประวัติการณ์ธรรมชาติ
- (b) จำนวนของวิดีโอ : มากกว่า 1,000,000 วิดีโอ
- (c) ความยาวเฉลี่ยของแต่ละวิดีโอ : 3 วินาที
- (d) จำนวนของหมวดหมู่ : 339 หมวดหมู่
- (e) วิธีการเก็บรวบรวมข้อมูล
 - i. เริ่มจากการรวบรวมคำที่ใช้อยู่ทั่วไปในชีวิตประจำวันมา 4,500 คำจาก VerbNet^[11] เว็บไซต์ที่เก็บรวบรวมคำกริยาภาษาอังกฤษขนาดใหญ่ จากนั้นนำมาแบ่งกลุ่มคำที่มีความหมายใกล้เคียงกันโดยใช้คุณลักษณะจาก Propbank^[15] และ FrameNet^[2] โดยเก็บข้อมูลเป็นแบบเวกเตอร์คุณลักษณะฐานสอง (binary feature vector) ซึ่งถ้าคำใดมีความเกี่ยวข้องกับคุณลักษณะก็จะให้ค่าเป็น 1 ถ้าไม่เกี่ยวข้องกันจะให้ค่าเป็น 0 จากนั้นจึงใช้วิธี k-means clustering ในการแบ่งกลุ่ม เมื่อแบ่งกลุ่มแล้วจำนวนจะเลือกคำจากในแต่ละกลุ่มนั้น โดยคำที่เลือกมาจะเป็นคำที่ใช้บ่อยที่สุดในกลุ่มนั้น และลบคำนั้นออกจากกลุ่มอื่นๆทั้งหมด (คำที่หนึ่งสามารถอยู่ได้หลายกลุ่ม) จากนั้นจะทำการวนการนี้ไปเรื่อยๆ แต่คำที่เลือกมาจะต้องไม่มีความหมายคลุมเครือไม่สามารถมองเห็นหรือได้ยินได้ และต้องไม่มีความหมายเหมือนกับคำที่เคยเลือกมา ก่อน จนสุดท้ายแล้วได้ออกมาที่ 339 หมวดหมู่
 - ii. ต่อมาทำการหาชุดข้อมูลวิดีโอด้วยจะตัดออกมาเพียง 3 วินาทีที่เกี่ยวข้องกับคำใน 339 หมวดหมู่ที่เลือกมาจากวิดีโอแหล่งต่างกัน 10 แหล่ง การตัดวิดีโอนั้นจะไม่ใช้พวก Video2Gif (โมเดลที่ระบุตำแหน่งของสิ่งที่มีส่วนในวิดีโอ) เพราะจะทำให้เกิดอคติขึ้นจะเกิดขึ้นตอนสร้างโมเดลจากนั้นจะทำการส่งข้อมูลของคำ และวิดีโอที่ตัดไปยัง Amazon Mechanical Turk (AMT หรือตลาดแรงงาน) เพื่อทำการสร้างคำกำกับโดยพนักงานของ AMT ทำให้ได้ 64 วิดีโอที่เกี่ยวข้องกับคำนั้น และอีก 10 วิดีโอที่มีคำกำกับอยู่แล้ว โดยวิดีโอที่มีคำกำกับอยู่แล้วนั้นถ้าพนักงานของ AMT ตอบเหมือนกันกิน 90% ถึงจะนำเข้าไปรวมกับชุดข้อมูลส่วนอีก 64 วิดีโอ ถ้าเป็นชุดข้อมูลสำหรับสร้างโมเดลจะต้องผ่านพนักงานของ AMT อย่างน้อย 3 ครั้ง และต้องมีคำกำกับเหมือนกัน 75% ขึ้นไปถึงจะถือว่าเป็นคำกำกับที่ถูกต้อง ถ้าเป็นชุดข้อมูลสำหรับตรวจสอบ และชุดข้อมูลสำหรับทดสอบจะต้องผ่านพนักงานของ AMT อย่างน้อย 4 ครั้ง และต้องมีคำกำกับเหมือนกัน 85% ขึ้นไป เหตุผลที่ไม่ตั้งเกณฑ์ไว้ที่ 100% เพราะจะทำให้วิดีโอนั้นยากเกินไปที่จะทำให้สามารถจำกัดการทำงานได้

2. การเตรียมข้อมูล

- (a) ชุดข้อมูลสำหรับสร้างโมเดลจะมี 802,264 วิดีโอ และมีวิดีโoinแต่ละหมวดหมู่อยู่ที่ 500 ถึง 5,000 วิดีโอ
- (b) ชุดข้อมูลสำหรับตรวจสอบจะมี 33,900 วิดีโอ และมีวิดีโoinแต่ละหมวดหมู่อยู่ที่ 100 วิดีโอ
- (c) แยกเฟรม RGB ออกมาจากวิดีโอ และทำการเปลี่ยนขนาดให้เป็น 340x256 pixel
- (d) ใช้อัลกอริทึม TVL1 optical flow จาก OpenCV เพื่อคัดข้อมูลrgbกวนที่จะเกิดขึ้น
- (e) ทำการแปลงค่าที่อยู่ใน optical flow ให้เป็นเลขจำนวนเต็มเพื่อทำให้การคำนวณนั้นเร็วขึ้น

- (f) ปรับค่า displacement ใน optical flow ให้ค่าสูงสุดเป็น 15 ต่ำสุดเป็น 0 และทำการปรับขนาดให้เป็นช่วง 0 - 255
- (g) เก็บข้อมูลอุปกรณ์ในรูปแบบของภาพขาวดำเพื่อลดพื้นที่ในการเก็บข้อมูล
- (h) แก้ปัญหาเรื่องการเคลื่อนไหวของกล้องด้วยการนำค่าเฉลี่ยของเวกเตอร์ไปลบกับ displacement
- (i) สุดท้ายจะเป็นสุ่มตัดภาพอุปกรณ์เพื่อเพิ่มจำนวนข้อมูล

3. โมเดลปัญญาประดิษฐ์

- (a) ในงานวิจัยนี้มีการทดสอบโมเดลปัญญาประดิษฐ์หลายรูปแบบ โดยโมเดลปัญญาประดิษฐ์ที่มีประสิทธิภาพการทำงานที่ดีที่สุด 5 ลำดับแรกมีดังนี้
 - i. SVM มีรูปแบบข้อมูลที่ป้อนเข้า คือ เพرمที่ต่อเนื่อง (spatial) + เพرمเดี่ยว (temporal) + ข้อมูลเสียง (auditory)
 - ii. I3D มีรูปแบบข้อมูลที่ป้อนเข้า คือ เพرمที่ต่อเนื่อง + เพرمเดี่ยว
 - iii. TRN-Multiscale มีรูปแบบข้อมูลป้อนเข้า คือ เพرمที่ต่อเนื่อง + เพرمเดี่ยว
 - iv. TSN-2stream มีรูปแบบข้อมูลป้อนเข้า คือ เพرمที่ต่อเนื่อง + เพرمเดี่ยว
 - v. ResNet50-ImageNet มีรูปแบบข้อมูลป้อนเข้า คือ เพرمที่ต่อเนื่อง
- (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 แล้วนำทั้ง 2 โมเดลไปทดสอบกับชุดข้อมูลอื่น โดยจะปรับอัตราความถี่ของเพرم (frame rate) ของวิดีโอให้เป็น 5 fps

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

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

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

4. ปัญหาที่พบ

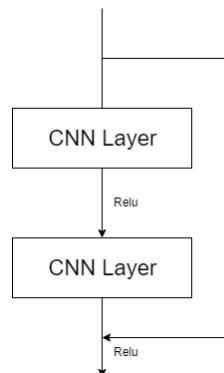
ผลลัพธ์จากการทำนายด้วยโมเดลถ้าผ่านรูปภาพที่มีรายละเอียดเยอะจะทำให้การทำนายโอกาสผิดนั้นค่อนข้างสูง ซึ่งปัญหานี้สามารถทำให้เกิดน้อยลงด้วยการนำวิธี class activation mapping (CAM) จะเป็นการเน้นรูปภาพในส่วนที่มีข้อมูลมากที่สุดและทำนายผลออกมานั้นๆ แต่ก็ยังมีจุดที่เป็นปัญหาอยู่ เช่น การกระทำที่เกิดขึ้นเร็วมากจะทำให้การทำนายนั้นมีโอกาสผิดสูงขึ้น

2.2 เครื่องมือสำหรับการวิเคราะห์ผลวิดีโอ

2.2.1 โมเดลปัญญาประดิษฐ์สำหรับจำแนกการกระทำมนุษย์

2.2.1.1 ResNet

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



รูปที่ 2.2: หลักการของ Residual block ของ ResNet

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

| จำนวนชั้นของ | Training error | |
|--------------|----------------|--------|
| | plain | ResNet |
| 18 | 27.94 | 27.88 |
| 34 | 28.54 | 25.03 |

ตารางที่ 2.6: อัตราเร้อยละของความผิดพลาดของชุดข้อมูลทดสอบ ImageNet

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

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

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

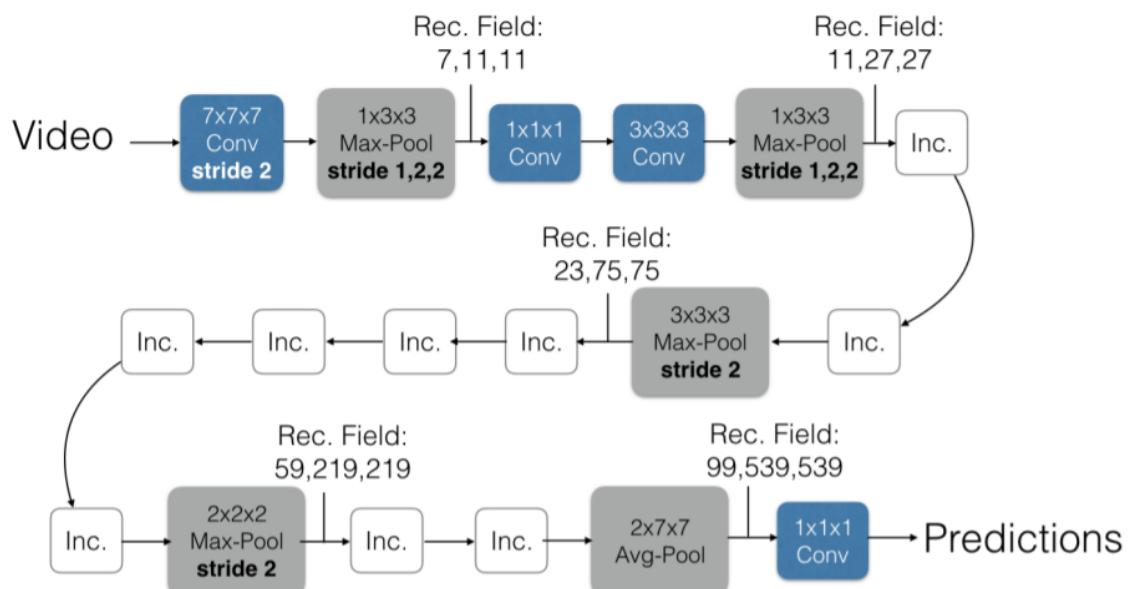
ตารางที่ 2.7: ค่าความผิดพลาดที่ได้จากการทดลองจำนวนชั้นของโมเดลปัญญาประดิษฐ์ ResNet บนชุดของข้อมูล CIFAR-10

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

2.2.1.2 Inflated 3D convolutional network

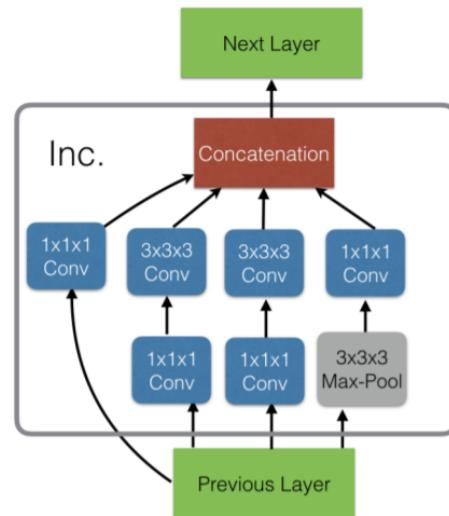
ในการพัฒนาโมเดลปัญญาประดิษฐ์สำหรับจำแนกการกระทำของมนุษย์นั้นมีพื้นฐานจากการจำแนกวัตถุ (object classification) หมายถึงการใช้รูปภาพหนึ่งรูปในการประมวลผลและคำนวณอุกมาดาว่าภายในรูปนั้นมีบริบทการกระทำอย่างไร โดยไม่ได้คำนึงถึงข้อมูลเชิงต่อเนื่อง (spatio-temporal information) จากบทความ ”Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset”^[4] นั้นได้พัฒนาโครงสร้างของโมเดลปัญญาประดิษฐ์ (architecture) ที่มีประสิทธิภาพในการประมวลผลภาพเคลื่อนไหวได้ชื่อว่า I3D หรือ inflated 3D-convolution network โดยโครงสร้างพื้นฐานของ I3D นั้นมาจากการนำ Inception-v1^[12] ที่ถูกพัฒนาโดย Google ซึ่งเป็นโครงสร้างที่มีประสิทธิภาพสูงในการจำแนกวัตถุในรูปภาพ และ I3D นั้นได้ทำการขยายมิติของโครงสร้างจาก 2 มิติ เป็น 3 มิติ เพื่อให้โมเดลปัญญาประดิษฐ์สามารถเรียนรู้ข้อมูลเชิงต่อเนื่องได้

Inflated Inception-V1



รูปที่ 2.3: โครงสร้างของโมเดลปัญญาประดิษฐ์ I3D^[4]

Inception Module (Inc.)



รูปที่ 2.4: โครงสร้างของโมเดลปัญญาประดิษฐ์ I3D^[4]

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

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

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

2.2.2 เครื่องมือสำหรับสร้างขุดข้อมูล

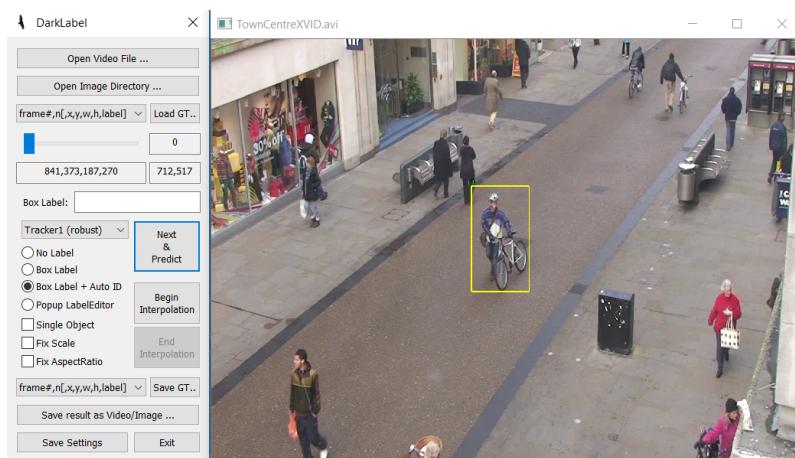
จากการค้นคว้าหาเครื่องมือในการสร้างคำกำกับข้อมูลเพื่อใช้เป็นแนวทางในการออกแบบเครื่องมือสำหรับกำกับข้อมูลด้วยปัญญาประดิษฐ์ พบรเครื่องมือที่เปิดให้ใช้งานสาธารณะ (open source) 2 เครื่องมือ คือ DarkLabel และ OpenLabeling โดยสรุปข้อสำคัญได้ดังนี้

โปรแกรม DarkLabel

เป็นโปรแกรมที่ช่วยในการทำนายคำกำกับและบันทึกในรูปแบบต่างๆ รองรับข้อมูลป้อนเข้าในรูปแบบไฟล์วิดีโอ avi, mp4 หรือกลุ่มรูปภาพ มีขั้นตอนการสร้างคำกำกับดังนี้

1. สร้างกรอบสี่เหลี่ยมครอบบริเวณวัตถุที่สนใจโดยใช้มนุษย์เป็นคนสร้าง
2. กดปุ่ม Next และ Predict อย่างต่อเนื่อง เพื่อทำนายตำแหน่งต่อไปของกรอบสี่เหลี่ยมในเฟรมถัดๆไป จนกระทั่งการเกิดข้อผิดพลาด
3. ลบกรอบสี่เหลี่ยมที่พลาด และเริ่มทำขั้นตอนที่ 1 ใหม่ อีกรังสีจันครบทุกเฟรมในวิดีโอ

หลังจากที่ผู้ใช้ได้ทดลองใช้โปรแกรม DarkLabel พบว่า เป็นโปรแกรมที่การทำงานส่วนใหญ่เป็นการสร้างคำกำกับแบบใช้มนุษย์เป็นคนทำด้วยตัวเอง ซึ่งทำให้ใช้เวลาในการทำงาน



รูปที่ 2.5: UI ของโปรแกรม DarkLabel

โปรแกรม OpenLabeling

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

1. การทำงานแบบอัตโนมัติ

หลังจากป้อนวิดีโอเข้าไปในโปรแกรมแล้วมีขั้นตอนการสร้างคำจำกัดดังนี้

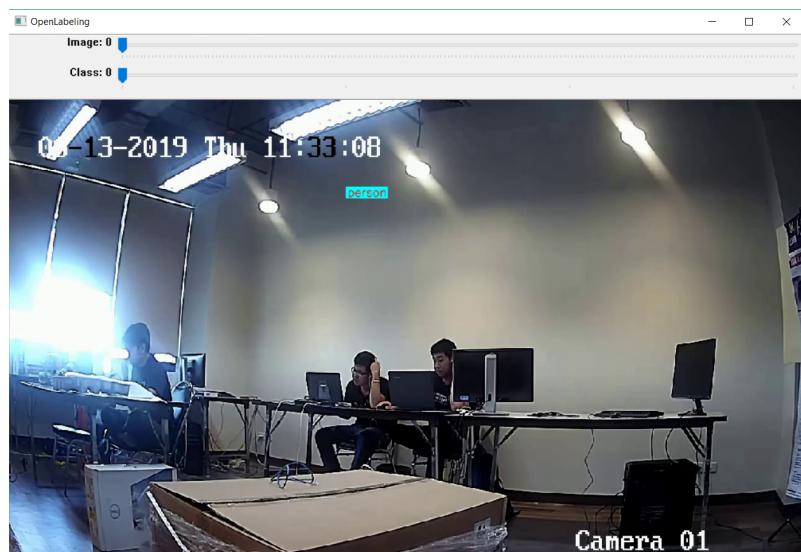
- (a) โปรแกรมจะทำงานอัตโนมัติโดยใช้โมเดลปัญญาประดิษฐ์ในการทำนายคีย์เฟรม (keyframe) และทำนายตำแหน่งต่อไปของกรอบสี่เหลี่ยมในเฟรมถัดไปด้วยอัลกอริทึมที่ใช้การคำนวนคณิตศาสตร์และการประมวลผลภาพในภาพที่เหลือ ผลลัพธ์ที่ได้คือรูปภาพและไฟล์คำจำกัดกับคุณลักษณะ

2. การทำงานแบบทำด้วยตัวเอง

หลังจากป้อนวิดีโอเข้าไปในโปรแกรมแล้วมีขั้นตอนการสร้างคำจำกัดดังนี้

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

หลังจากที่ได้ทดลองใช้โปรแกรม OpenLabeling ทั้ง 2 รูปแบบการทำงานแล้วพบว่า การทำงานแบบอัตโนมัติ ไม่สามารถปรับแก้ไขสิ่งใดในระหว่างกระบวนการนั้น ทำให้หากเกิดกรณีที่โมเดลทำนายกรอบสี่เหลี่ยมพลาดหรือเกินมา จะไม่สามารถแก้ไขได้ และการทำงานแบบทำด้วยตัวเองไม่มีระบบตรวจสอบกรอบสี่เหลี่ยม ทำให้ผู้ใช้งานจะต้องสร้างกรอบสี่เหลี่ยมขึ้นมาเอง

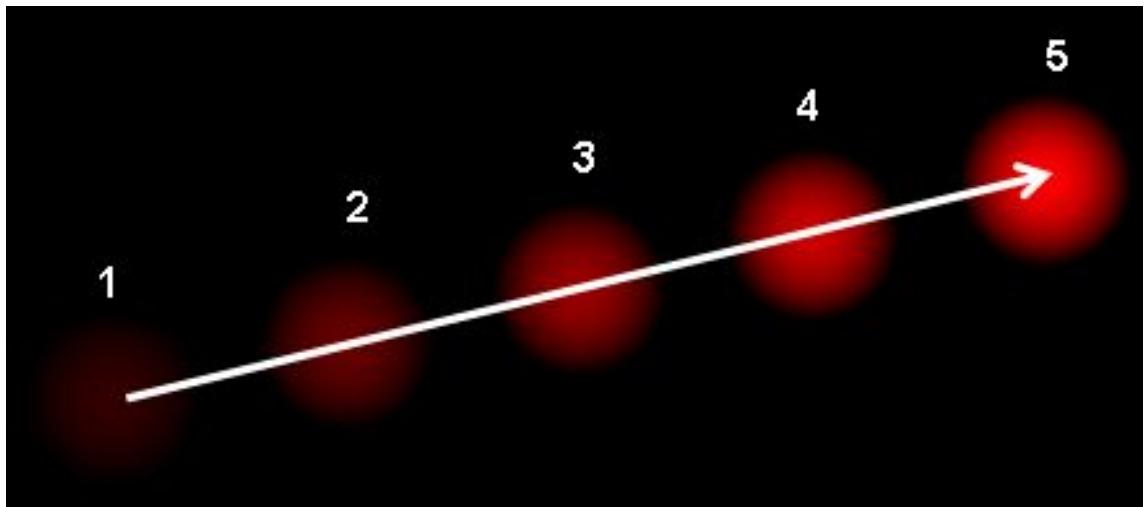


รูปที่ 2.6: UI ของโปรแกรม OpenLabeling

2.3 ทฤษฎีเกี่ยวข้อง

2.3.1 Optical flow

Optical flow^[1] คือการแปลงการเคลื่อนที่ของวัตถุในระหว่างสองรูปภาพซึ่งอาจจากการจากเคลื่อนที่ของวัตถุหรือตัวกล้องออกมารูปแบบของเวกเตอร์ 2 มิติ โดยที่เวกเตอร์แต่ละตัวจะแสดงถึงทิศทางการเคลื่อนที่ระหว่างภาพดังรูปด้านล่าง



รูปที่ 2.7: ตัวอย่างการเคลื่อนที่ของลูกบอล

จากรูปภาพจะแสดงให้เห็นถึงการเคลื่อนที่ของลูกบอลของภาพที่ต่อเนื่องกัน 5 ภาพโดยที่ลูกคระแสดงถึงทิศทางการเคลื่อนที่ของเวกเตอร์

การทำงานของ optical flow อยู่บนสมมติฐานหลายประการได้แก่

- ความเข้มพิกเซล (pixel) ของวัตถุจะไม่เปลี่ยนแปลงระหว่างภาพที่ต่อเนื่องกัน
- พิกเซลที่อยู่ใกล้กันจะมีการเคลื่อนไหวที่คล้ายกัน

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

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

โดยที่

I = พิกเซลจากภายในภาพ

x = ตำแหน่งของพิกเซลในแกน x

dx = ระยะทางที่เคลื่อนที่ในแกน x

y = ตำแหน่งของพิกเซลในแกน y

dy = ระยะทางที่เคลื่อนที่ในแกน y

t = เวลา

dt = ระยะเวลาที่เปลี่ยนไประหว่างภาพ

จากนั้นใช้การประมาณค่าของ taylor series ทางฝั่งขวาเมื่อและลบค่า common term แล้วหารด้วย dt เพื่อให้ได้สมการดังต่อไปนี้

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

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

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

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 โดยเป็นเกรเดียนของภาพ และ f_t เป็นเกรเดียนของเวลา แต่ n กับ b เป็นตัวแปรที่ไม่ทราบ ทำให้สมการนี้ไม่สามารถแก้ไขโดยมีตัวแปรที่ไม่ทราบถึง 2 ตัว จึงมีการนำวิธีการต่างๆเข้ามาใช้ในการแก้ปัญหานี้ โดยวิธีการที่นำเข้ามาใช้ในการแก้ปัญหาก็คือ dense optical flow ซึ่งใช้อัลกอริทึมของ Gunner Farneback^[6] ซึ่งจะใช้วิธีการขยายพหุนาม (polynomial expansion)

บทที่ 3

ระเบียบวิธีวิจัย

ในการทำโครงการวิจัยเครื่องมือสำหรับกำกับข้อมูลด้วยปัญญาประดิษฐ์ จะมีการทำงานหลากหลายส่วนมาทำงานร่วมกัน ซึ่งต้องมีระเบียบวิธีวิจัยอธิบายถึงขั้นตอนการดำเนินงานตั้งแต่เริ่มศึกษาข้อมูลจนไปถึงสิ้นสุดกระบวนการวิจัย โดยใช้ภาษาไทย เป็นภาษาหลักในการเขียนโปรแกรม

3.1 ความต้องการของระบบ

3.1.1 ความต้องการเชิงการใช้งาน (functional requirements)

1. เครื่องมือสำหรับกำกับข้อมูลด้วยปัญญาประดิษฐ์ต้องสามารถตัดวิดีโอช่วงเวลาที่ไม่มีมนุษย์อยู่ออกได้อัตโนมัติโดยใช้ปัญญาประดิษฐ์
2. เครื่องมือสำหรับกำกับข้อมูลด้วยปัญญาประดิษฐ์สามารถระบุตำแหน่งมนุษย์แต่ละคนในวิดีโอด้วยการกระทำของมนุษย์ในวิดีโอได้ โดยการกระทำที่กำหนดจะประกอบไปด้วย ยืน นั่ง นอน เล่นโทรศัพท์ เดิน กินข้าว พูดคุย
3. ชุดข้อมูลที่ได้จากเครื่องมือสำหรับกำกับข้อมูลด้วยปัญญาประดิษฐ์ต้องสามารถนำไปใช้ในการพัฒนาโมเดลปัญญาประดิษฐ์ต่อได้
4. สร้างระบบต้นแบบของเครื่องมือสำหรับกำกับข้อมูลด้วยปัญญาประดิษฐ์ที่มีมนุษย์สามารถทำงานร่วมกับปัญญาประดิษฐ์ได้
5. ระบบวิเคราะห์การกระทำมนุษย์ต้องสามารถนำวิดีโอมาวิเคราะห์ข้อมูลการกระทำและตำแหน่งของมนุษย์แต่ละคน และนำข้อมูลเหล่านี้ไปสร้างรายงานออกมากได้ โดยรายละเอียดรายงานจะมีดังนี้
 - (a) เวลา (time stamp)
 - (b) รหัสระบุตัวตน (ID)
 - (c) การกระทำ
 - (d) ตำแหน่ง โดยจะบอกในลักษณะของกรอบสี่เหลี่ยมครอบพื้นที่ที่มีมนุษย์คนนั้นๆอยู่

3.1.2 ความต้องการเชิงวิศวกรรม (non-functional requirements)

1. สร้างเครื่องมือสำหรับกำกับข้อมูลด้วยปัญญาประดิษฐ์โดยใช้ภาษาไทย
2. ความละเอียดอย่างต่ำของวิดีโอต้องมากกว่า 640 x 480 (กว้าง x สูง)
3. วิดีโожะต้องมีอัตราเฟรมต่อวินาที (fps) อย่างต่ำ 10 เฟรมต่อวินาที

3.2 หน้าที่ความรับผิดชอบ

ปฐมพงศ์ สินธุ์งาม สร้างและทดสอบโมเดลปัญญาประดิษฐ์สำหรับจัดทำการกระทำมนุษย์ 3D รวมถึงออกแบบและสร้างระบบ Tracker

ศุภกร เบญจวิกรัย รวมฟังก์ชันและระบบต่างๆของแอพพลิเคชัน รวมถึงออกแบบและสร้างระบบ Select และ Detect

อุกฤษฎ์ เลิศวรรณาการ สร้างและทดสอบโมเดลปัญญาประดิษฐ์สำหรับจัดทำการกระทำมนุษย์ Resnet-50 รวมถึงออกแบบและสร้างระบบ Person ReID

3.3 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

3.4 ภาษาที่ใช้ในการพัฒนาระบบ

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

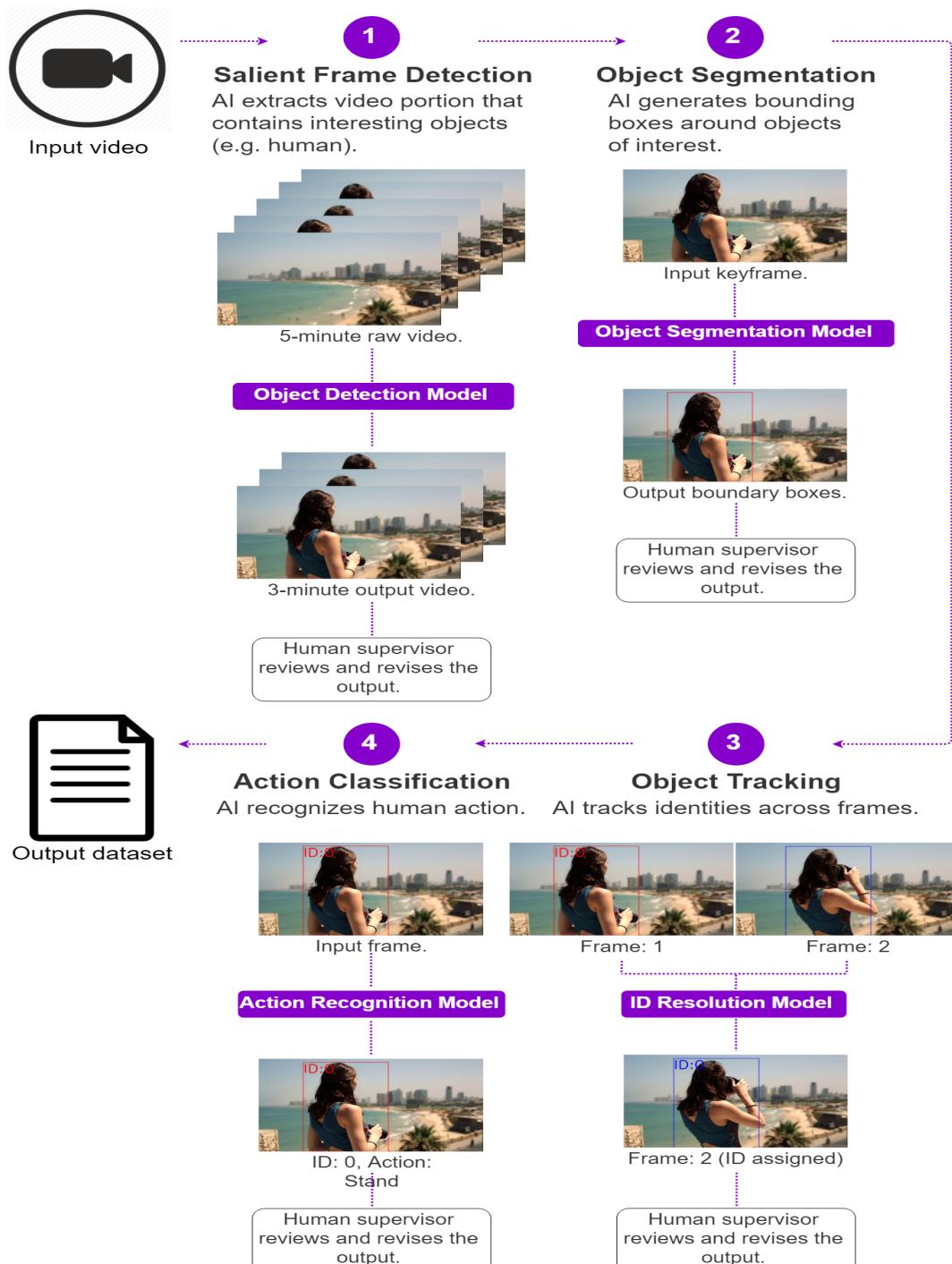
3.5 Program library ที่ใช้ในการพัฒนาระบบและแอปพลิเคชัน

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

3.6 แผนการดำเนินงาน

โดยจากที่กล่าวไปตอนต้นในบทนำการดำเนินงานและการออกแบบการสร้างเครื่องมือสำหรับกำกับข้อมูล ด้วยปัญญาประดิษฐ์ และระบบบิเคราะห์การกระทำการของมนุษย์ในวิดีโอ มีแผนการทำงานซึ่งถูกแบ่งออกเป็นสาม ขั้นตอนดังนี้ ขั้นตอนแรกคือ ขั้นตอนของการศึกษาทำความเป็นไปได้ รวมถึงเทคโนโลยีปัจจุบันที่เกี่ยวกับการ สร้างแอปพลิเคชัน และการจัดการกระทำการของมนุษย์ด้วยปัญญาประดิษฐ์ เพื่อนำมาประยุกต์ใช้กับงานวิจัย นี้ ขั้นตอนที่สองคือ ขั้นตอนของการออกแบบและสร้างแอปพลิเคชันที่ใช้ในการสร้างชุดข้อมูลสำหรับการเทรน ไมเดลจากวิดีโอ ขั้นตอนที่สามคือ ขั้นตอนของการออกแบบและสร้างระบบบิเคราะห์การกระทำการของมนุษย์ได้โดย มีข้อกำหนดตามที่กล่าวไว้ในบทนำ ในการเริ่มทำงานวิจัยนี้นั้นสิ่งจำเป็นที่ต้องทำในอันดับแรกคือการศึกษาข้อมูล ในหัวข้อที่เกี่ยวข้อง หรืองานวิจัยอื่นที่ทำเอาระบบแล้ว เพื่อศึกษาและทำความเข้าใจ ข้อดี-ข้อเสีย ของเทคนิคหรือ กระบวนการต่างๆ เพื่อนำมาประยุกต์ใช้กับงานวิจัยนี้ ในการศึกษาเกี่ยวกับการออกแบบและ การสร้างแอปพลิ เคชันที่ใช้ในการสร้างชุดข้อมูลสำหรับการสร้างโมเดลจากวิดีโอ สิ่งที่ต้องให้ความสนใจคือฟังก์ชันการทำงาน การ ออกแบบและการจัดวางองค์ประกอบต่างๆในหน้าต่างแอปพลิเคชัน และความสะดวกในการใช้งาน จากนั้นจึงเริ่ม ศึกษาเกี่ยวกับ library ที่ใช้ในการสร้างแอปพลิเคชัน ส่วนการศึกษาเกี่ยวกับการสร้างระบบบิเคราะห์การกระทำ มนุษย์ จะมุ่งความสนใจไปที่ชุดข้อมูลสำหรับการวิเคราะห์วิดีโอ ไมเดลสำหรับการวิเคราะห์วิดีโอ เทคนิคในการ สร้างโมเดล เทคโนโลยีในการระบบบิเคราะห์วิดีโอ เพื่อใช้ในการออกแบบและสร้างระบบบิเคราะห์การกระทำ ของมนุษย์ในวิดีโอด้วยมีประสิทธิภาพ ในบทนี้จะกล่าวถึงกระบวนการออกแบบและการดำเนินการตามแผนที่วางแผนไว้

3.7 ภาพรวมระบบของเครื่องมือสำหรับกำกับข้อมูลด้วยปัญญาประดิษฐ์



รูปที่ 3.1: ภาพรวมระบบของเครื่องมือสำหรับกำกับข้อมูลด้วยปัญญาประดิษฐ์

3.8 การออกแบบหน้าต่างแอพพลิเคชันของเครื่องมือสำหรับกำกับข้อมูลด้วยปัญญาประดิษฐ์

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

3.8.1 เครื่องมือสำหรับกำกับข้อมูลด้วยปัญญาประดิษฐ์

แอพพลิเคชันแบ่งการทำงานออกเป็นสี่ส่วนประกอบด้วยกระบวนการ Select, Detect, Track และ Label เพื่อช่วยแบ่งเบ้าภาระของผู้พัฒนาในการสร้าง label สำหรับสร้างโมเดลจากข้อมูลประเภทวิดีโอ โดยกระบวนการ Select จะต้องสามารถตัดวิดีโอ่วันที่ไม่มีมนุษย์อยู่ออกจากวิดีโอด้วย กระบวนการ Detect จะต้องหาตำแหน่งของมนุษย์ภายในวิดีโอด้วย แล้วใช้กระบวนการ Track นำรายตำแหน่งต่อไปของมนุษย์ข้อมูลตำแหน่งของมนุษย์ที่ได้จากการกระบวนการ Detect และกระบวนการ Label นั้นต้องสามารถทำงานร่วมกับปัญญาประดิษฐ์ได้ ดังรูปที่ 3.2

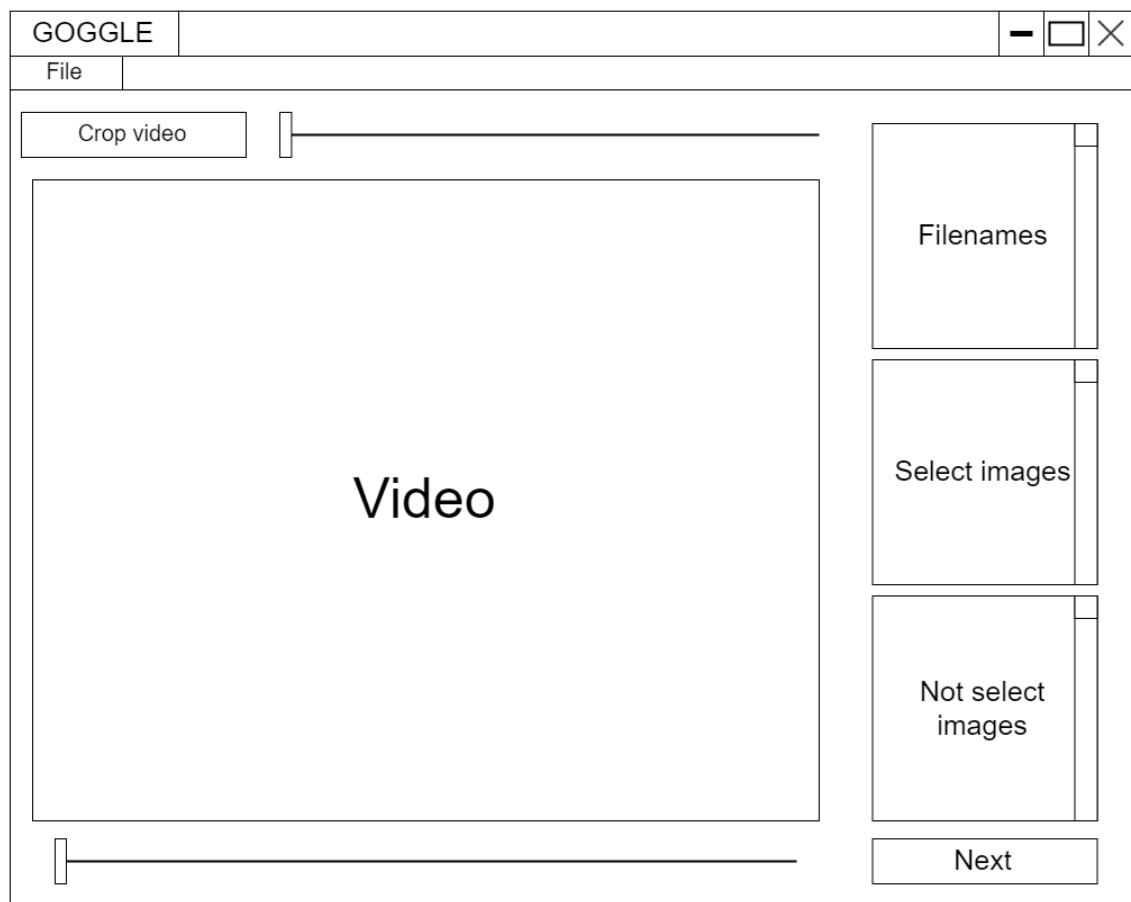


รูปที่ 3.2: กระบวนการหลักของเครื่องมือสำหรับกำกับข้อมูลด้วยปัญญาประดิษฐ์

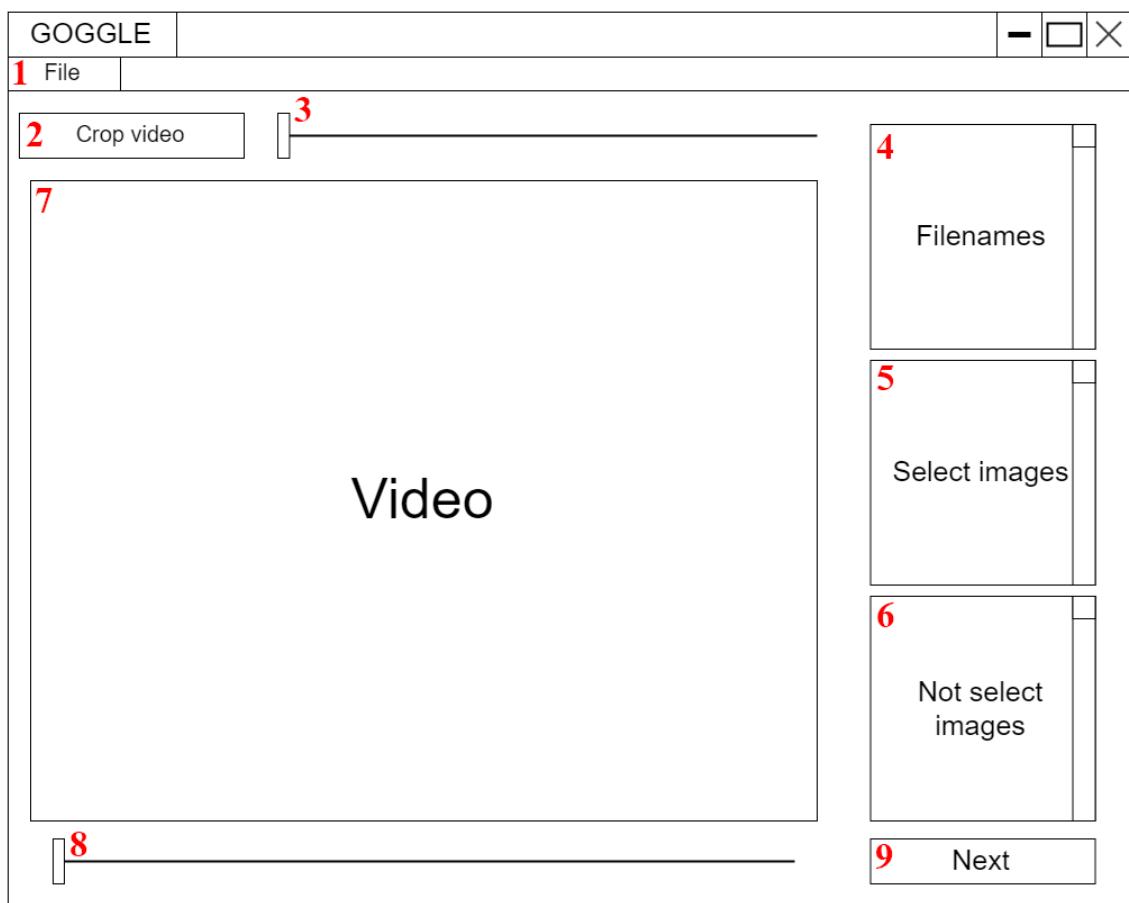
โดยแต่ละกระบวนการจะมีรายละเอียดดังนี้

3.8.1.1 Select

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



รูปที่ 3.3: หน้าต่าง Select ของเครื่องมือสำหรับกำกับข้อมูลด้วยปัญญาประดิษฐ์



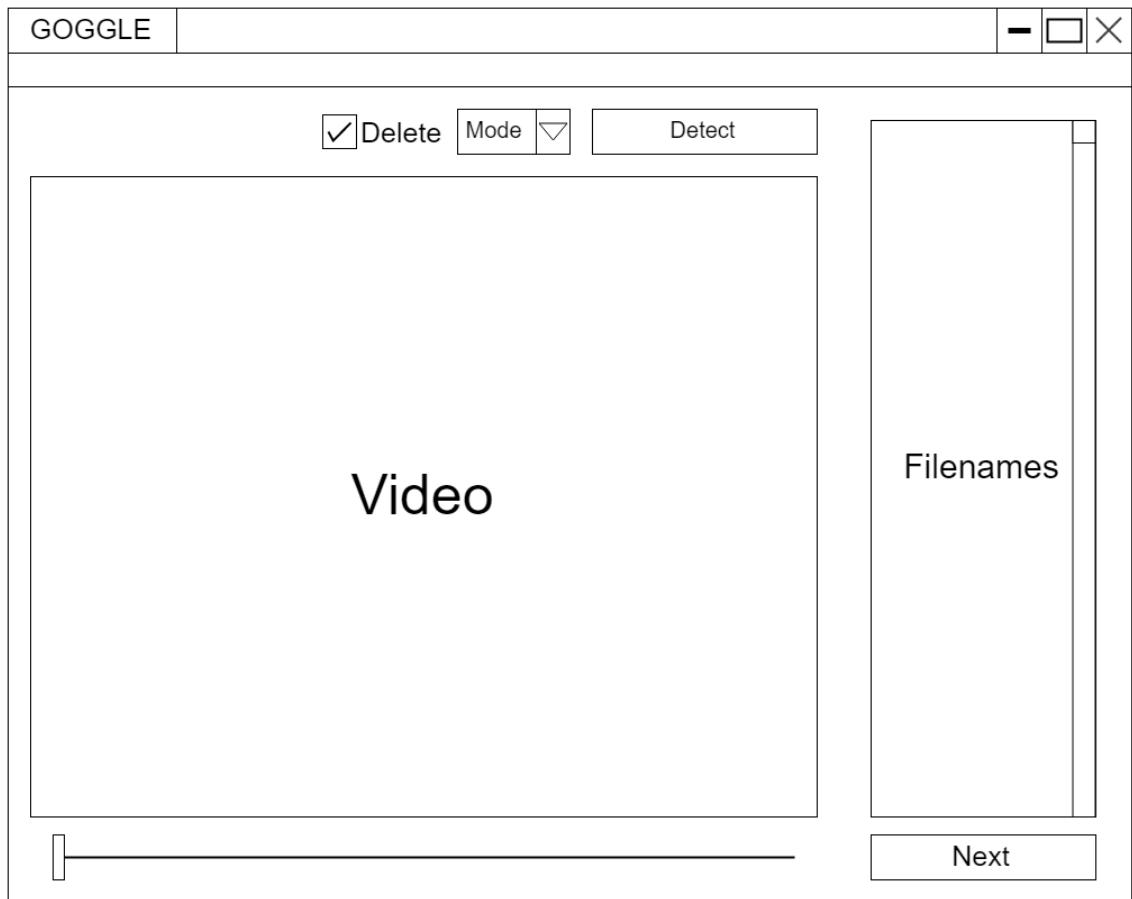
รูปที่ 3.4: ตำแหน่งของแต่ละวิดเจ็ตในหน้าต่าง Select

โดยที่แต่ละวิดเจ็ตตามหมายเลขที่กำหนดตามรูปที่ 3.4 มีรายละเอียดดังนี้

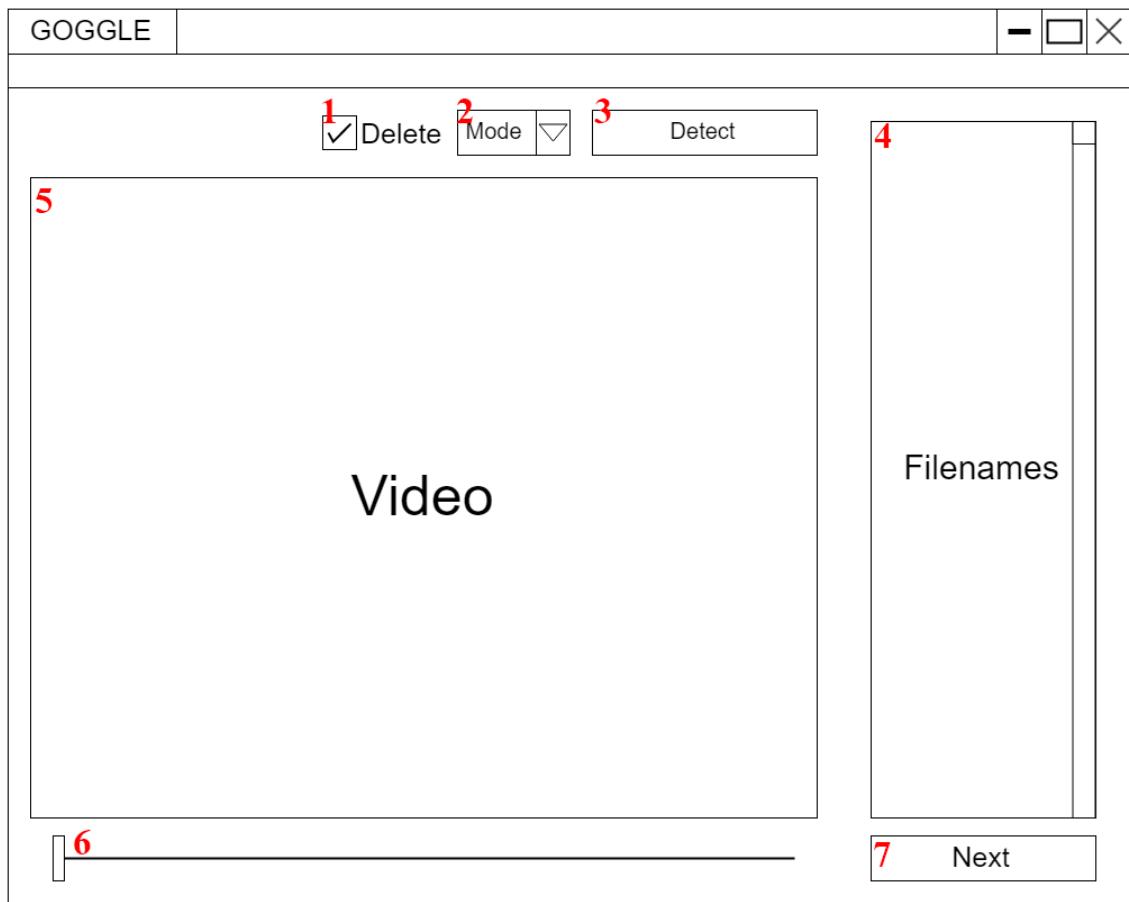
1. หมายเลข 1 คือปุ่มสำหรับเลือกไฟล์วิดีโอที่ต้องการจากในคอมพิวเตอร์เข้ามาในโปรแกรม
2. หมายเลข 2 คือปุ่มสำหรับสั่งให้ระบบทำการสร้างคีย์เฟรมขึ้นมา แล้วใช้ปัญญาประดิษฐ์ประมวลผลเพื่อแยกคีย์เฟรมในเมื่อน้อย และคีย์เฟรมในเมื่อน้อยแบบอัตโนมัติ (Auto mode)
3. หมายเลข 3 คือแถบเลื่อนเพื่อกำหนดความถี่ในการหยิบคีย์เฟรม 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 จะต้องสามารถรับคีย์เฟรมจากกระบวนการ Select มาประมวลผลด้วยปัญญาประดิษฐ์เพื่อหาตำแหน่งของมนุษย์ที่อยู่ในคีย์เฟรม แล้วสร้างกรอบสีเหลี่ยมครอบบริเวณดังกล่าวได้ในแบบอัตโนมัติ เพื่อแบ่งเบาภาระผู้ใช้ในการที่ต้องสร้างกรอบสีเหลี่ยมครอบตำแหน่งของมนุษย์ด้วยตัวเอง และผู้ใช้ต้องสามารถสร้างหรือลบกรอบสีเหลี่ยมได้ด้วยตัวเองสำหรับแก้ไขความผิดพลาดของปัญญาประดิษฐ์ เพื่อเพิ่มคุณภาพของชุดข้อมูล จึงออกแบบหน้าต่างได้ดังรูปที่ 3.5



รูปที่ 3.5: หน้าต่าง Detect ของเครื่องมือสำหรับกำกับข้อมูลด้วยปัญญาประดิษฐ์



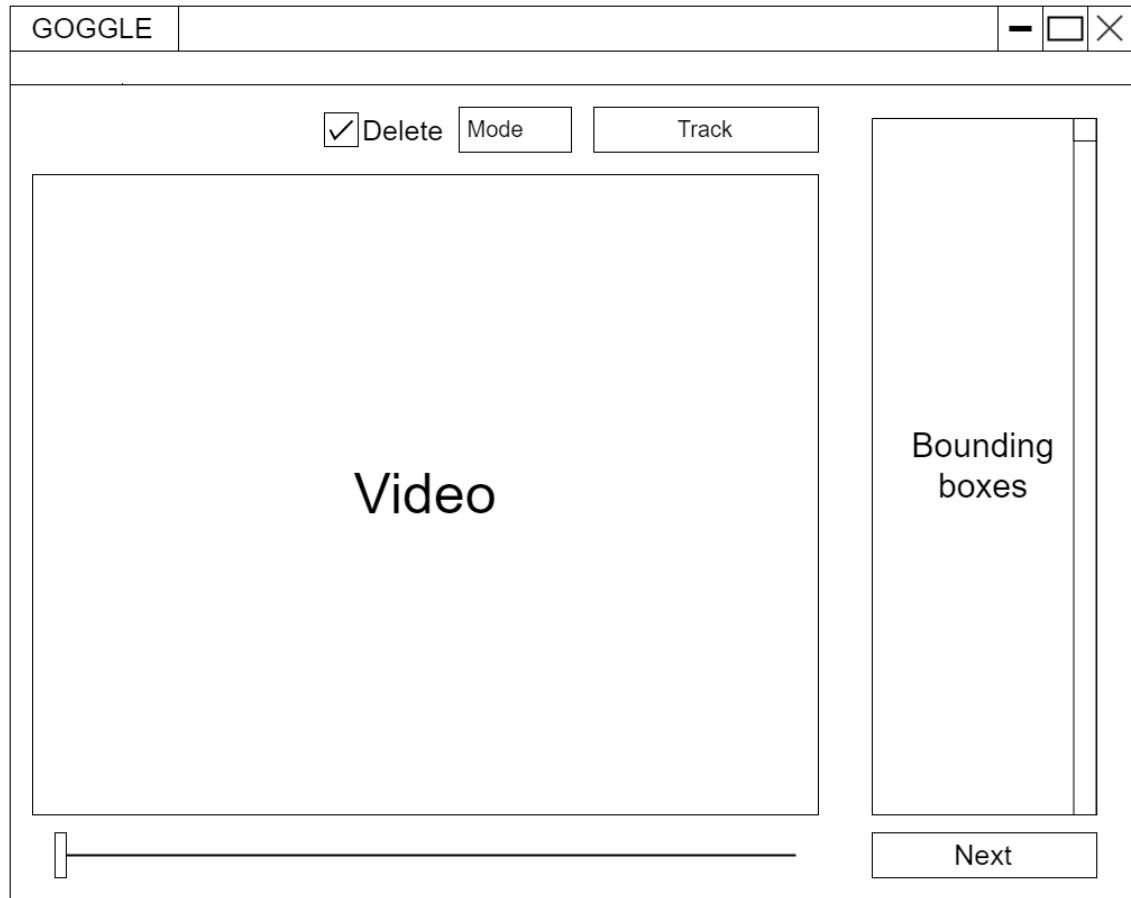
รูปที่ 3.6: ตำแหน่งของแต่ละวิดเจ็ตในหน้าต่าง Detect

โดยที่แต่ละวิดเจ็ตตามหมายเลขที่กำหนดตามรูปที่ 3.6 มีรายละเอียดดังนี้

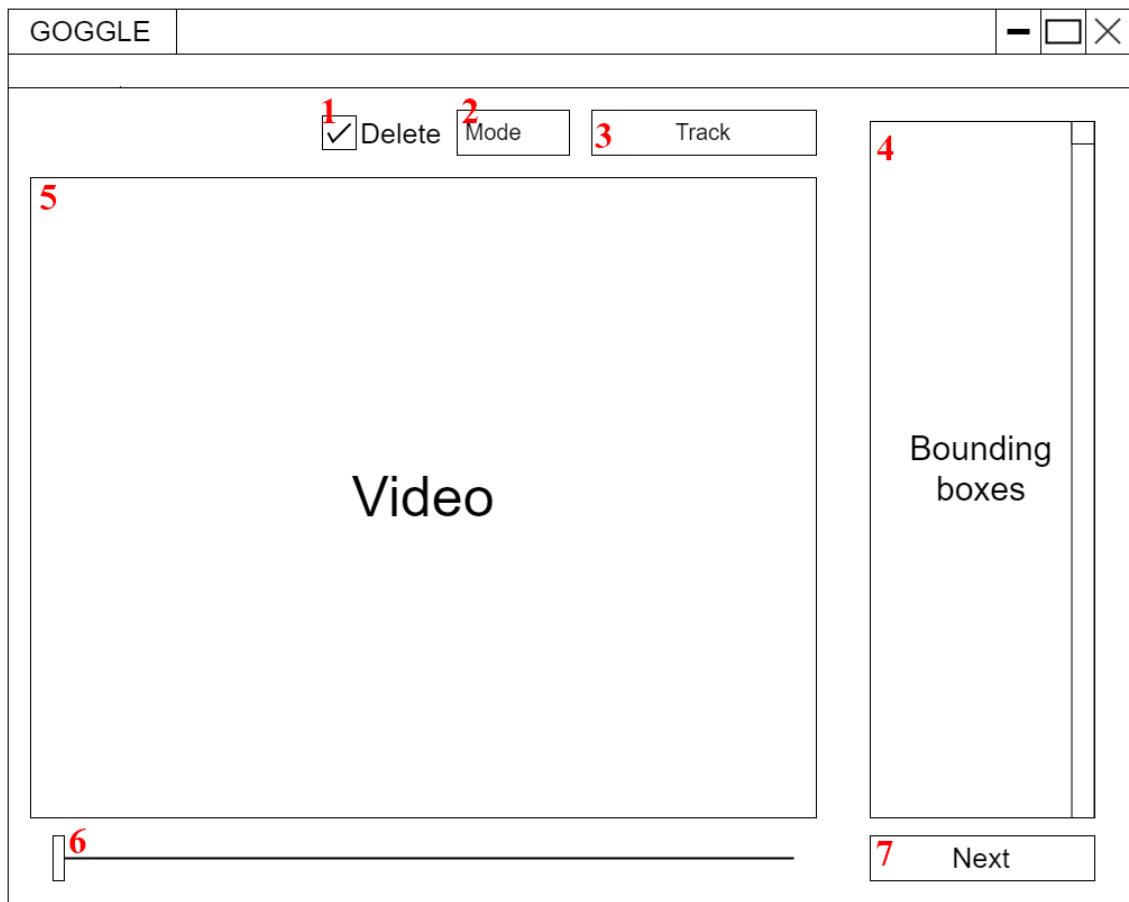
1. หมายเลข 1 คือช่องสำหรับกดเพื่อเปลี่ยนระบบจากสร้างกรอบสี่เหลี่ยมในแบบแก้ไขด้วยตนเอง (Manual mode) เป็นลบรอบสี่เหลี่ยมแทน
2. หมายเลข 2 คือช่องสำหรับเลือกว่าจะใช้ระบบแบบใด ระหว่างแบบอัตโนมัติและแบบแก้ไขด้วยตนเอง
3. หมายเลข 3 คือปุ่มสำหรับสั่งให้ระบบทำการตรวจหาตำแหน่งของมนุษย์ในคิร์เฟรมทั้งหมดแล้วสร้างกรอบสี่เหลี่ยมขึ้นมาครอบบริเวณที่กำหนด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



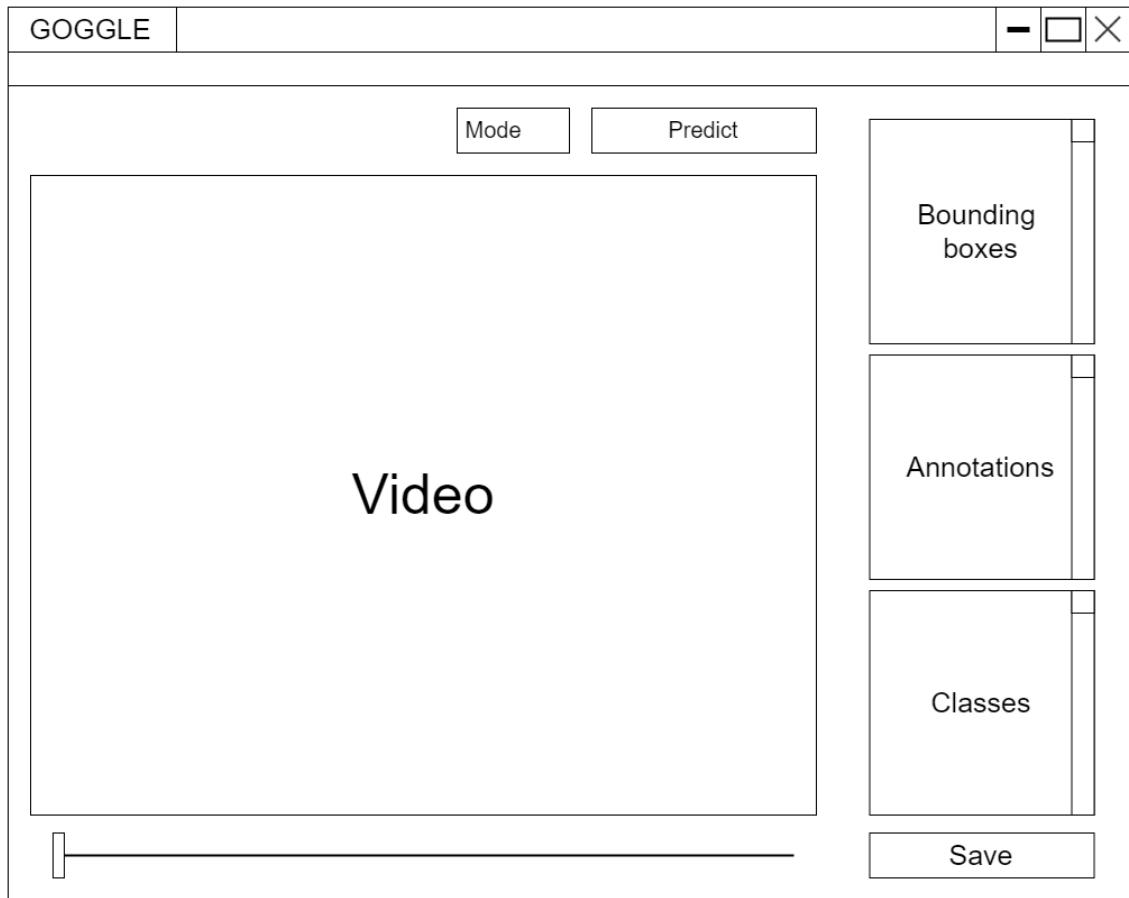
รูปที่ 3.8: ตำแหน่งของแต่ละวิดเจ็ตในหน้าต่าง Track

โดยที่แต่ละวิดเจ็ตตามหมายเลขที่กำหนดตามรูปที่ 3.8 มีรายละเอียดดังนี้

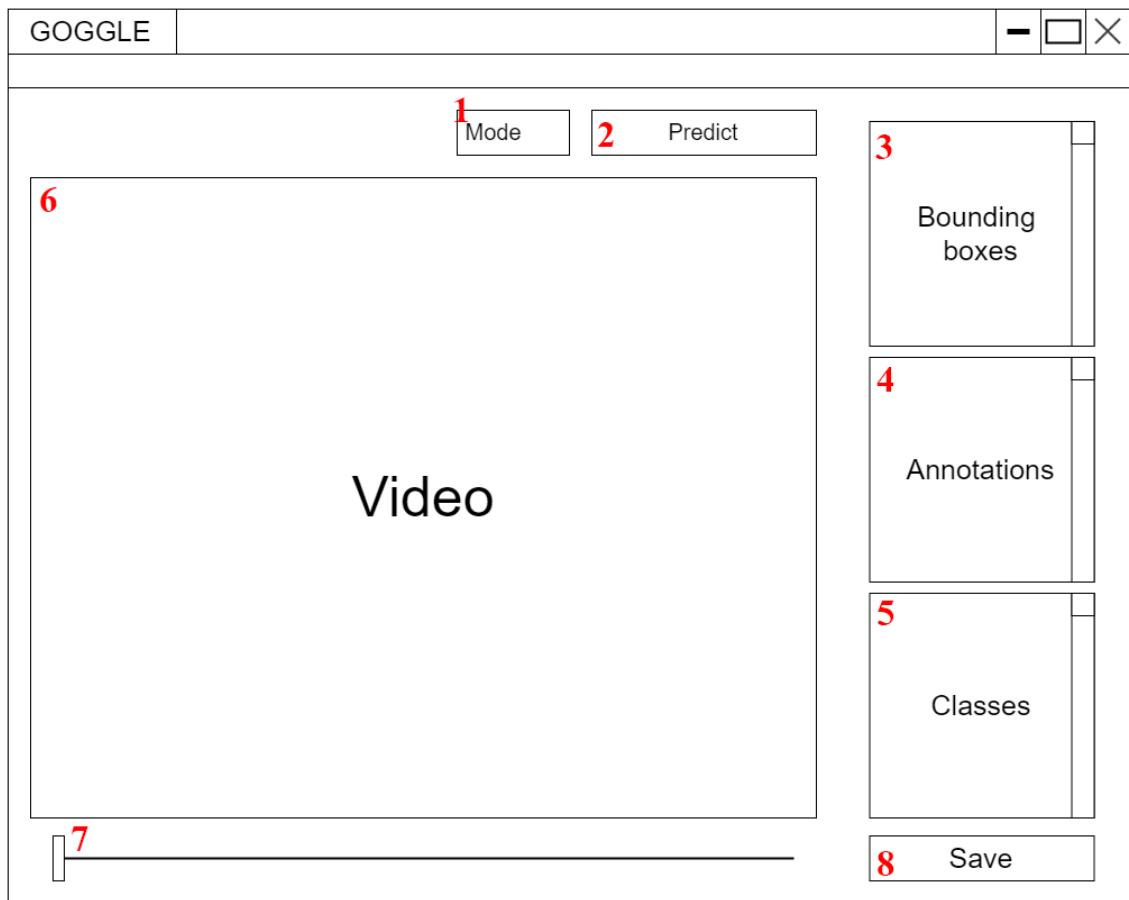
1. หมายเลข 1 คือช่องสำหรับกดเพื่อเปลี่ยนระบบจากสร้างกรอบสี่เหลี่ยมในแบบแก้ไขด้วยตนเองเป็นลบกรอบสี่เหลี่ยมแทน
2. หมายเลข 2 คือช่องสำหรับเลือกว่าจะใช้ระบบแบบใด ระหว่างแบบอัตโนมัติและแบบแก้ไขด้วยตนเอง
3. หมายเลข 3 คือปุ่มสำหรับสั่งให้ระบบทำการตรวจสอบตำแหน่งของมนุษย์ในเฟรมระหว่างคิ้ยวเฟรมทั้งหมดแล้วสร้างกรอบสี่เหลี่ยมขึ้นมาครอบบริเวณที่กำหนด
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 คือปุ่มสำหรับสั่งให้ระบบคำนวณรูปแบบที่มีอยู่ในทุกๆเฟรม
3. หมายเลข 3 คือกล่องสำหรับแสดงกรอบสี่เหลี่ยมทั้งหมดที่มีอยู่ในเฟรมที่เลือก
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 นั้นเป็นรูปแบบที่นิยมใช้ในการเก็บข้อมูลสำหรับการสร้างโมเดลประเพณีตรวจสอบจับตุ๊ก โดยจะเก็บข้อมูลในรูปแบบของ PASCAL VOC ที่นิยมใช้ในการสร้างโมเดลด้วย library Tensorflow โดยภายในไฟล์จะมีข้อมูลดังรูปที่ 3.11 โดยข้อมูลส่วนสำคัญของรูปแบบนี้นั้นจะถูกใส่หมายเลขอ้างอิงแต่ละหมายเลขนั้นหมาย

```

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

รูปที่ 3.11: ตัวอย่างข้อมูลภายในไฟล์ xml

ถึง

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

3.9 การออกแบบการทดสอบการตรวจจับวัตถุ

3.9.1 ทดสอบประสิทธิภาพการทำงานของโมเดลปัญญาประดิษฐ์สำหรับการทำการตรวจจับภาพบุคคล สิ่งที่ใช้ในการวัดผล

1. ความเร็วต่อรูปภาพ (วินาที)
2. ความแม่นยำ โดยคำนึงถึงอัตราส่วนร่วมของกรอบที่เหลือ หรือ Intersection over Union (IoU)

จุดประสงค์

1. ผู้วิจัยได้ตั้งจุดประสงค์การทดลอง การใช้โมเดลปัญญาประดิษฐ์สำหรับตรวจจับวัตถุ เพื่อวัดผลโมเดล ปัญญาประดิษฐ์ที่ใช้ในปัจจุบัน และหาโมเดลปัญญาประดิษฐ์สำหรับการทำการตรวจจับวัตถุที่มีความเร็วมาก ที่สุดและแม่นยำสูงที่สุดเมื่อทดสอบกับชุดข้อมูลของผู้วิจัย

ตัวแปร

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

ตัวแปรควบคุม

1. ชุดข้อมูล : ชุดข้อมูลสำหรับทดสอบวัดผลที่ผู้วิจัยสร้างขึ้น (สูม 20 เฟรมจากวิดีโอที่ผู้วิจัยใช้สำหรับสร้าง ชุดข้อมูล)

วิธีการทดลอง

1. แบ่งชุดข้อมูลออกเป็นชุดข้อมูลสำหรับทดสอบ และชุดข้อมูลที่มีคำจำกัดเพื่อเป็นคำตอบ
 - (a) ชุดข้อมูลสำหรับทดสอบ ประกอบด้วย : ชื่อของวิดีโอ และเฟรม
 - (b) ชุดข้อมูลที่มีคำจำกัดเพื่อเป็นคำตอบ ประกอบด้วย : ชื่อของวิดีโอ เฟรม และตำแหน่งของกรอบ สีเหลี่ยม
2. เรียกชื่อและเฟรมของวิดีโอด้วยชุดข้อมูลทดสอบ และนำโมเดลปัญญาประดิษฐ์ที่มีคำจำกัดเพื่อเป็นคำตอบ เก็บผลลัพธ์เป็นชุดข้อมูลผลลัพธ์จากการทำนาย

 - (a) ชุดข้อมูลผลลัพธ์จากการทำนาย ประกอบด้วย : ชื่อของวิดีโอ เฟรม และตำแหน่งของกรอบ สีเหลี่ยม

3. ประเมินผลการทำงานโดยเทียบระหว่างชุดผลลัพธ์จากการทำนาย และชุดข้อมูลที่มีคำจำกัดเพื่อเป็นคำตอบ ผ่านฟังก์ชันคำนวนค่า IoU
4. เปรียบเทียบผลลัพธ์จากแหล่งที่มา

3.10 การออกแบบการทดสอบการทำงานตามตำแหน่งต่อไปของมุนช์

3.10.1 ทดสอบประสิทธิภาพการทำงานของระบบทำงานตำแหน่งต่อไปของวัตถุในวิดีโอ สิ่งที่ใช้ในการวัดผล

1. ความเร็วต่อวิดีโอ (วินาที)
2. ความแม่นยำ โดยคำนึงถึงอัตราส่วนร่วมของกรอบที่เหลือ

สมมุติฐาน

ผู้วิจัยได้ตั้งสมมุติฐานว่า การใช้โมเดลปัญญาประดิษฐ์สำหรับตรวจจับวัตถุและสร้างกรอบสี่เหลี่ยมทุกๆ N เฟรม แล้วใช้ระบบทำงานตำแหน่งต่อไปของวัตถุในการสร้างกรอบสี่เหลี่ยมในเฟรมระหว่างนั้น จะทำให้ระบบสามารถทำงานได้เร็วขึ้น โดยที่ประสิทธิภาพจะลดลงเพียงเล็กน้อย

ตัวแปรควบคุม

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

วิธีการทดลอง

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

3.11 การออกแบบการทดสอบการระบุตัวตนของบุคคลภายในภาพ

3.11.1 ทดสอบประสิทธิภาพการทำงานของระบบระบุตัวตนของบุคคลภายในภาพ

สิ่งที่ใช้ในการวัดผล

- ความแม่นยำสำหรับการระบุตัวตนของบุคคลภายในภาพ

สมมุติฐาน

ผู้จัดได้ตั้งสมมุติฐานว่า ผลลัพธ์ของการทดลองการใช้งานจริงของโมเดลปัญญาประดิษฐ์ ResNet50 ที่สร้างด้วยชุดข้อมูล Market1501 นั้นควรจะมีความแม่นยำในการระบุตัวตนของบุคคลภายในภาพมากที่สุดเมื่อเทียบกับโมเดลปัญญาประดิษฐ์ที่สร้างด้วยชุดข้อมูลอื่นๆ 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 สำหรับการทำหนังของบุคคล

วิธีการทดลอง

- ดาวน์โหลดโมเดลปัญญาประดิษฐ์ที่ถูกสร้างด้วยชุดข้อมูลต่างๆ ได้แก่ Market1501, DukeMTMCReID, CUHK03 และ MSMT17
- นำชุดข้อมูลที่ผู้จัดสร้างขึ้นมาผ่านโมเดลปัญญาประดิษฐ์ YOLO-V3 320 เพื่อหาตำแหน่งของบุคคล
- นำโมเดลปัญญาประดิษฐ์แต่ละอันมาทดสอบความแม่นยำสำหรับการระบุตัวตนของบุคคลภายในภาพ ด้วยตำแหน่งของบุคคลที่ได้มาจากการขึ้นตอนก่อนหน้านี้
- ประเมินผลการทำงานโดยเทียบความแม่นยำสำหรับการระบุตัวตนของบุคคลภายในภาพของแต่ละโมเดล ปัญญาประดิษฐ์ เพื่อหาโมเดลปัญญาประดิษฐ์ที่ได้ผลลัพธ์ดีที่สุด

3.12 การออกแบบการทดสอบการจดจำการกระทำของมนุษย์

3.12.1 ทดสอบประสิทธิภาพการทำงานของโมเดลปัญญาประดิษฐ์ที่เคยถูกтренร์ผ่าน AVA โดยใช้ชุดข้อมูลของ AVA ในการทดสอบและเทียบผลลัพธ์กับแหล่งอ้างอิง

สิ่งที่ใช้ในการวัดผล

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

สมมุติฐาน

ผู้จัยได้ตั้งสมมุติฐานว่า ผลลัพธ์ของการทดลองจะมีความแม่นยำเทียบเท่ากับผลลัพธ์จากแหล่งที่มา แต่ความเร็วต่อรูปภาพจะมีความเร็วน้อยกว่าผลลัพธ์จากแหล่งที่มา เนื่องจากแหล่งที่มาของข้อมูลได้ทำการทดสอบโดยใช้กราฟิกการ์ดรุ่น Nvidia GeForce GTX TITAN X ซึ่งเป็นกราฟิกการ์ดที่มีประสิทธิภาพการทำงานดีกว่ากราฟิกการ์ดของผู้จัย จึงทำให้สามารถทดสอบด้วยความเร็วที่มากกว่า

ตัวแปรควบคุม

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

วิธีการทดลอง

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

3.12.2 ทดสอบประสิทธิภาพการทำงานของโมเดลปัญญาประดิษฐ์ที่เคยถูกสร้างด้วย AVA และใช้ชุดข้อมูลที่ผู้วิจัยสร้างขึ้นในการทดสอบและเทียบผลลัพธ์กับแหล่งอ้างอิง สิ่งที่ใช้ในการวัดผล

1. ความเร็วต่อรูปภาพ (วินาที)
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. เรียกชื่อของวิดีโอจากชุดข้อมูลทดสอบ และนำโมเดลปัญญาประดิษฐ์ที่นายผลลัพธ์ จากนั้นเก็บผลลัพธ์ เป็นชุดข้อมูลผลลัพธ์จากการทำงาน
 - (a) ชุดข้อมูลผลลัพธ์จากการทำงาน ประกอบด้วย : ชื่อของวิดีโอ เพرم ตำแหน่งของกรอบสีเหลี่ยม รหัสของการกระทำ และความมั่นใจ
3. ประเมินผลการทำงานโดยเทียบระหว่างชุดผลลัพธ์จากการทำงาน และชุดข้อมูลที่มีคำกำกับเพื่อเป็นคำตอบ
4. เปรียบเทียบผลลัพธ์กับผลการทดลองที่ผ่านมา

3.12.3 ทดสอบประสิทธิภาพการทำงานของโมเดลปัญญาประดิษฐ์ที่ถูกสร้างด้วยชุดข้อมูลที่ผู้วิจัยสร้างขึ้น และใช้ชุดข้อมูลที่ผู้วิจัยสร้างขึ้นในการทดสอบและเทียบผลลัพธ์กับแหล่งอ้างอิง สิ่งที่ใช้ในการวัดผล

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

สมมุติฐาน

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

ตัวแปรควบคุม

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

วิธีการทดลอง

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

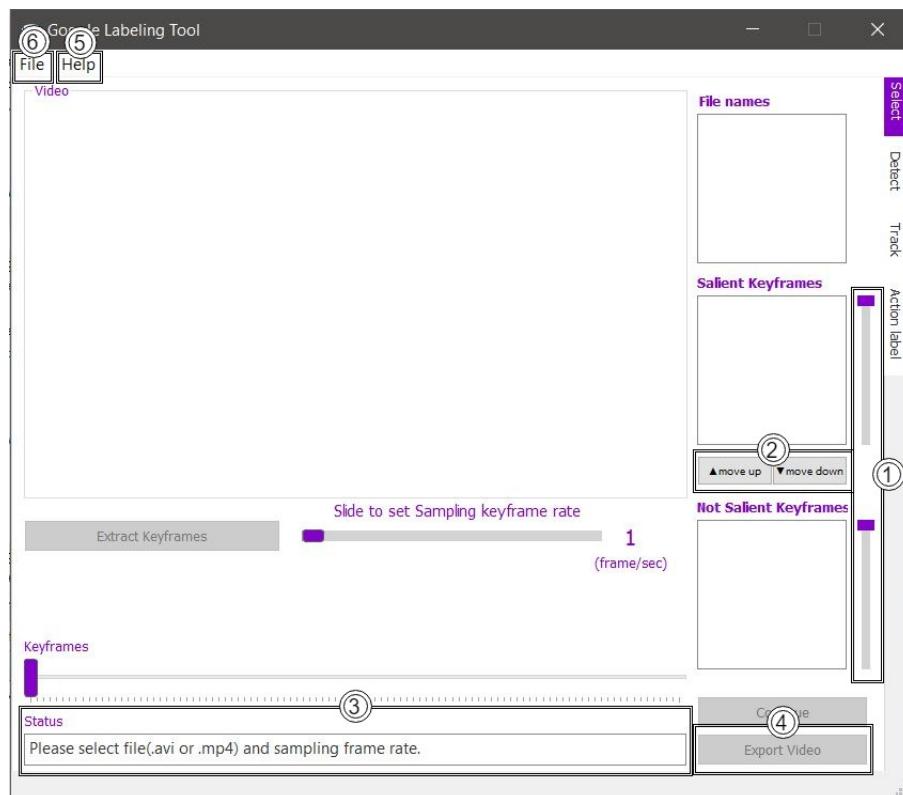
บทที่ 4

ผลการดำเนินงาน

4.1 Labeling tool

4.1.1 หน้าต่างแสดงผลของแอพพลิเคชัน

หน้าต่าง Select

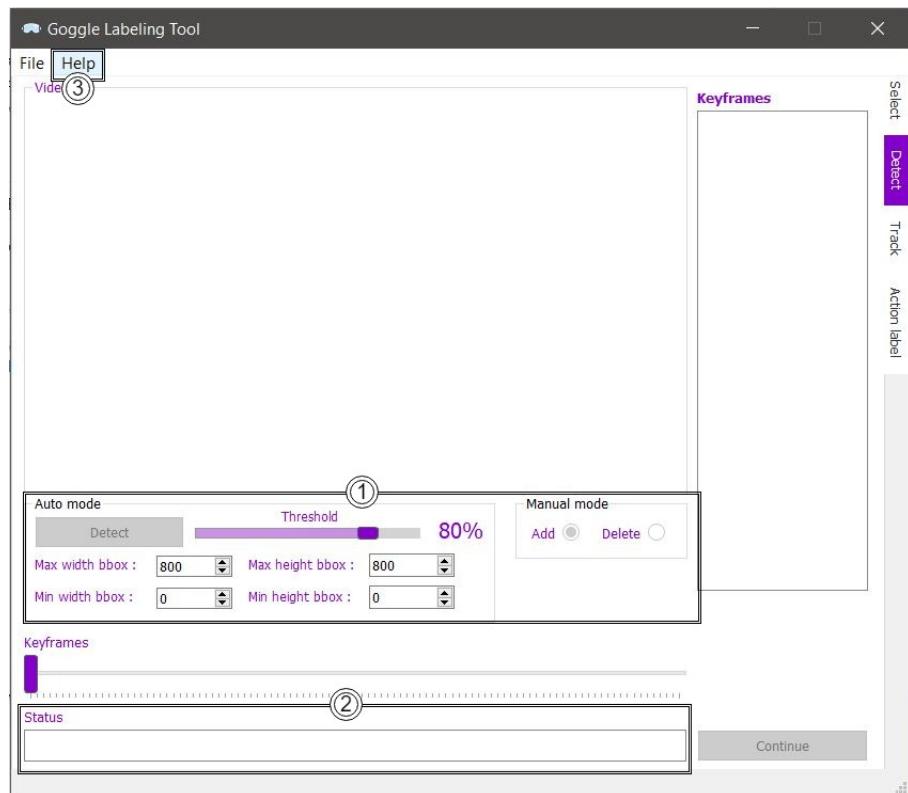


รูปที่ 4.1: รูปหน้าต่างแสดงผลของหน้าต่าง Select

จากรูปที่ 4.1 แสดงหน้าต่าง Select ของแอพพลิเคชัน ซึ่งเมื่อเทียบกับกับฉบับร่างตามรูปที่ 3.3 จะมีส่วนที่เพิ่มเติมขึ้นมาดังนี้

1. แถบเลื่อนสำหรับเลื่อนคุณูปกรณ์ที่มีมนุษย์หรือไม่มีมนุษย์ เพื่อเพิ่มความสะดวกในการเลือกคุณูปกรณ์
2. ปุ่มสำหรับแก้ไขคุณูปกรณ์ที่มีมนุษย์หรือไม่มีมนุษย์
3. แถบแสดงสถานะกระบวนการทำงาน
4. ปุ่มสำหรับนำผลลัพธ์ออกเป็นไฟล์วิดีโอด้วยไฟล์ในช่วงที่มีมนุษย์อยู่
5. แถบสำหรับคำแนะนำช่วยเหลือ
6. ปุ่มสำหรับเปิดไฟล์

หน้าต่าง Detect

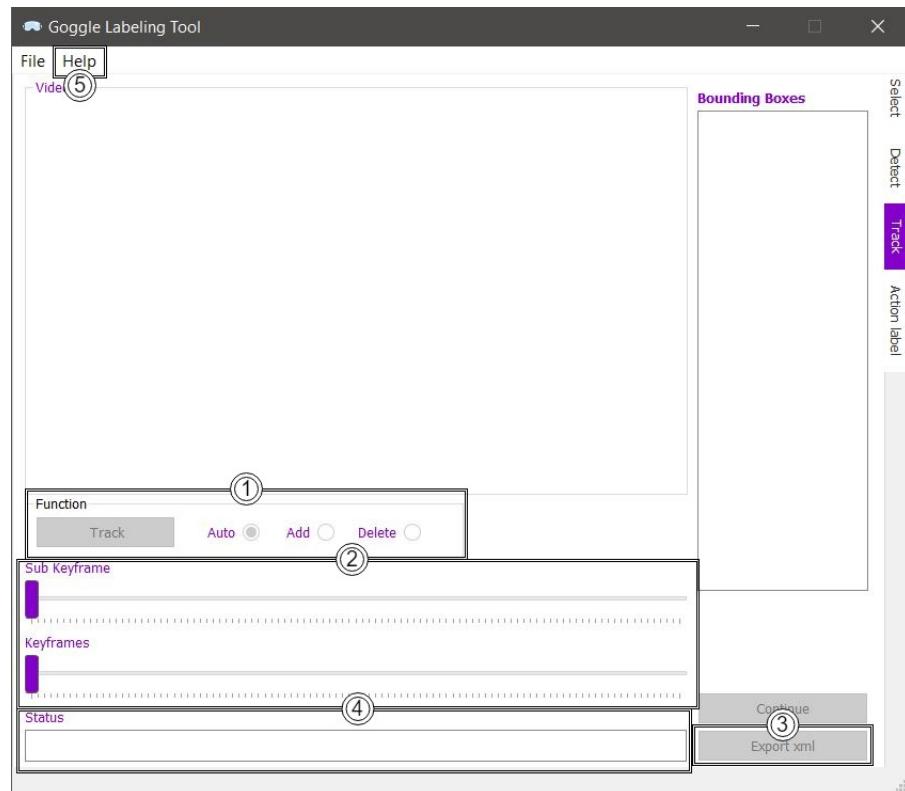


รูปที่ 4.2: รูปหน้าต่างแสดงผลของหน้าต่าง Detect

จากรูปที่ 4.2 แสดงหน้าต่าง Detect ของแอพพลิเคชัน ซึ่งเมื่อเทียบกับรูปที่ (3.5) จะมีส่วนที่เพิ่มเติมขึ้นมาดังนี้

1. ปรับหน้าตาใหม่ด้วยการทำงานแบบอัตโนมัติ และกำหนดเองสามารถใช้งานได้สะดวกขึ้น และเพิ่มความหลากหลายในการปรับแก้ในการทำงานอัตโนมัติ
2. แถบแสดงสถานะกระบวนการทำงาน
3. แถบสำหรับคำแนะนำช่วยเหลือ

หน้าต่าง Track

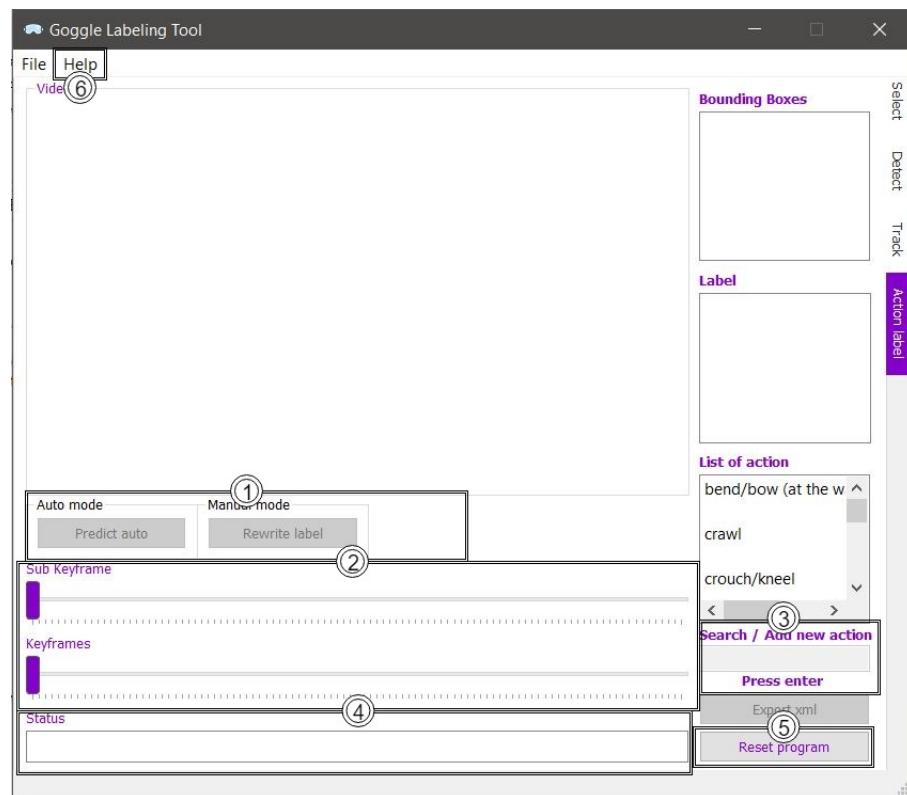


รูปที่ 4.3: รูปหน้าต่างแสดงผลของหน้าต่าง Track

จากรูปที่ 4.3 แสดงหน้าต่าง Track ของแอพพลิเคชัน ซึ่งเมื่อเทียบกับกับฉบับร่างตามรูปที่ (3.7) จะมีส่วนที่เพิ่มเติมขึ้นมาดังนี้

1. ปรับหน้าตาใหม่จากการทำงานแบบอัตโนมัติและกำหนดเองจากฉบับร่างเพื่อให้สามารถใช้งานได้สะดวกขึ้น
2. เพิ่มແລບເລືອນເປັນ 2 ແລບເລືອນທຳໃຫ້ສາມາຄຸດຄີ່ຍິ່ເຟຣີມແລະເຟຣີມທີ່ອຸ່ຽ່ຮ່ວ່າງໜ່ວງຄີ່ຍິ່ເຟຣີມໄດ້ສະດວກขື້ນ
3. เพິມປຸ່ມສໍາຫຼັບນຳພລັດພົບອອກເປັນໄຟລ໌ 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 ແຕບເລື່ອນທຳໃຫ້ສາມາດຄຸງຄິຍໝ່າງແລ້ວໄດ້ສະໜັກຂຶ້ນ
3. ເຄີຍອີ່ນສໍາຫຼັບຄັນຫາຫຼືເພີ່ມໝາວດໜູ້ຂອງການກະທຳ
4. ແຕບແສດງສະຖານະກະບວນການກະທຳ
5. ປຸ່ມສໍາຫຼັບເຮີມຕົ້ນການກະທຳໃໝ່
6. ແຕບສໍາຫຼັບຄຳແນະນຳໜ່ວຍເຫຼືອ

4.1.2 ผลลัพธ์การทำงานในแต่ละหน้าต่างของแอปพลิเคชัน

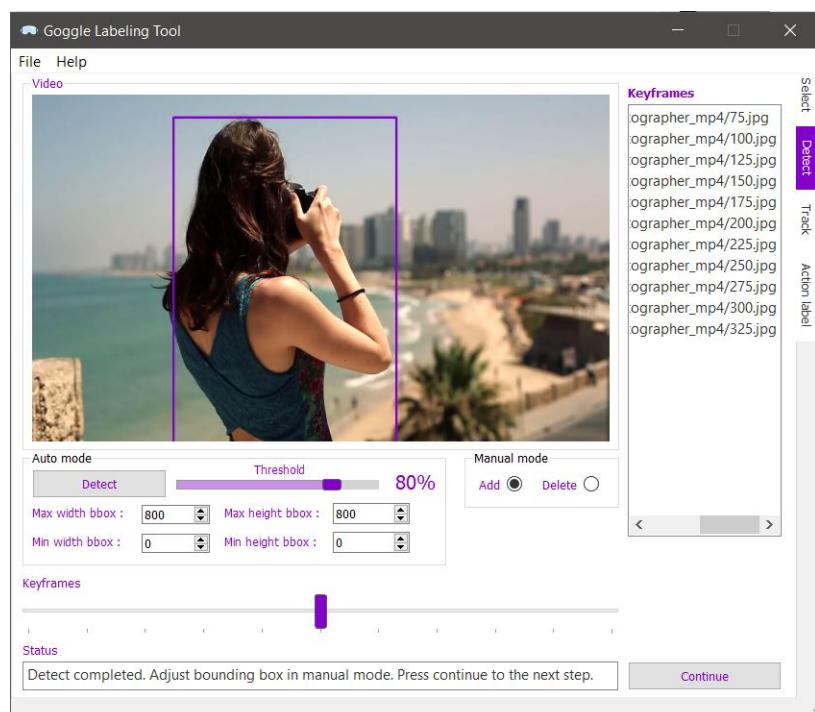
ผลลัพธ์การทำงานของหน้าต่าง Select



รูปที่ 4.5: รูปผลลัพธ์การแยกเฟรมที่มีมนุษย์อยู่ และไม่มีมนุษย์อยู่ภายในเฟรม

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

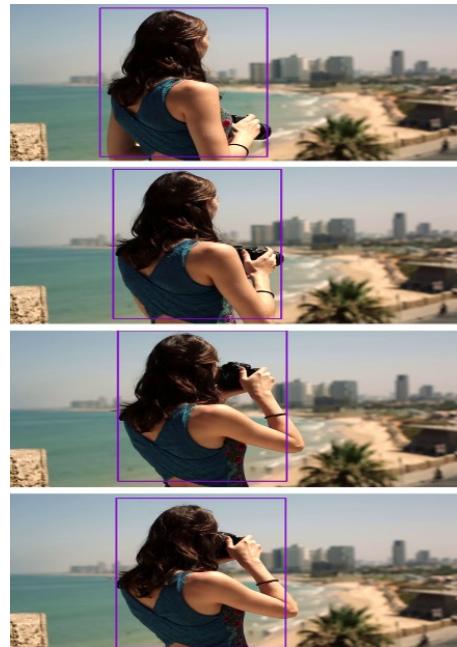
ผลลัพธ์การทำงานของหน้าต่าง Detect



รูปที่ 4.6: รูปคีย์เฟรมที่ถูกตีกรอบสีเหลืองในส่วนที่มีมนุษย์อยู่

แอ��พลิเคชันจะนำคีย์เฟรมที่มีนุชย์ที่ได้จากหน้าต่าง Select นำมาตีกรอบสี่เหลี่ยมในส่วนของเฟรมที่มีมนุชย์อยู่โดยสามารถใช้โหมดการทำงานแบบบอตโน้มัติหรือแบบแก้ไขเองก็ได้ ซึ่งผลลัพธ์ที่ได้จะได้คีย์เฟรมที่มีกรอบสี่เหลี่ยม ดังรูปที่ 4.6 จากนั้นจะบันทึกข้อมูลในไฟล์ txt

ผลลัพธ์การทำงานของหน้าต่าง Track



(ก) ตัวอย่างเฟรมที่ถูกตีกรอบสี่เหลี่ยม

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

(ข) ตัวอย่างไฟล์ xml

รูปที่ 4.7: รูปผลลัพธ์การทำงานของหน้าต่าง Track

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

ผลลัพธ์การทำงานของหน้าต่าง Label



(ก) ตัวอย่างเฟรมที่ถูกตีกรอบสีเหลืองและคำทำนายการกระทำ

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

(ข) ตัวอย่างไฟล์ xml

รูปที่ 4.8: รูปผลลัพธ์การทำงานของหน้าต่าง Label

แอพพลิเคชันจะนำกรอบสีเหลืองของทุกเฟรมที่มีมนุษย์อยู่มาทำนายมนุษย์ในกรอบสีเหลืองนั้นกำลังมีการกระทำการอะไรอยู่ โดยสามารถทำงานได้ทั้งหมดอัตโนมัติหรือแบบแก้ไขเอง และสามารถบันทึกข้อมูลออกเป็นไฟล์ xml ได้ดังรูปที่ 4.8x

4.2 ผลการทดลองการตรวจจับวัตถุ

4.2.1 ข้อมูลรายละเอียดประกอบการทดลอง

จำนวนเฟรมทั้งหมด: 20 เฟรม

จำนวนมนุษย์ที่อยู่ในเฟรม: 0-5 คน

ความละเอียดรูปภาพ: 1280×720 พิกเซล

ขอบเขตอัตราส่วนร่วมของกรอบที่เหลือที่จะนับว่าการทำนายถูกต้อง: 50% ขึ้นไป

4.2.2 ทดสอบประสิทธิภาพการทำงานของโมเดลปัญญาประดิษฐ์สำหรับการทำการตรวจจับภาพบุคคล

ข้อมูลผลการทำงานของโมเดลปัญญาประดิษฐ์สำหรับการทำการตรวจจับภาพบุคคล ข้างต้นเป็นข้อมูลจากเว็บไซต์ของ YOLO

| | ความเร็วต่อรูปภาพ (มิลลิวินาที) | ความแม่นยำ (0.5 IoU mAP) |
|--------------------------|---------------------------------|--------------------------|
| SSD Mobilenet v1 ppm | 26 | 20 |
| YOLO-v3 320 | 22 | 51.5 |
| YOLO-v3 tiny | 4.5 | 33.1 |
| YOLO-v3 spp | 50 | 60.6 |
| Faster RCNN inception v2 | 58 | 28 |

ข้อมูลผลการทำงานของโมเดลปัญญาประดิษฐ์สำหรับการทำการตรวจจับภาพบุคคลหลังจากการทดลอง

| | ความเร็วต่อรูปภาพ (มิลลิวินาที) | ความแม่นยำ (0.5 IoU mAP) |
|--------------------------|---------------------------------|--------------------------|
| SSD Mobilenet v1 ppm | 63 | 37 |
| YOLO-v3 320 | 65 | 64.9 |
| YOLO-v3 tiny | 17 | 44.4 |
| YOLOv3-spp | 65 | 70.3 |
| Faster RCNN inception v2 | 981 | 42.5 |

4.2.3 ทดสอบประสิทธิภาพการทำงานของโมเดลปัญญาประดิษฐ์สำหรับการทำการตรวจจับภาพบุคคล

ข้อมูลความแม่นยำของโมเดลปัญญาประดิษฐ์เมื่อทดสอบด้วยชุดข้อมูลของผู้วิจัย

| | ความเร็วต่อรูปภาพ(มิลลิวินาที) | ความแม่นยำ (0.5 IOU) |
|-------------------------|--------------------------------|----------------------|
| SSD Mobilenet v1 ppm | 63 | 37 |
| YOLOv3-320 | 65 | 64.9 |
| YOLOv3-tiny | 17 | 44.4 |
| YOLOv3-spp | 65.4 | 70.3 |
| Faster rcnn inceptionv2 | 981 | 42.5 |

ตารางที่ 4.1: ข้อมูลผลการทำงานของโมเดลปัญญาประดิษฐ์สำหรับการทำการตรวจจับภาพบุคคลหลังจากการทดลอง

4.3 ผลการทดสอบการทำนายตำแหน่งต่อไปของมนุษย์

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 ทดสอบประสิทธิภาพ และความเร็วในการประมวลผล

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

ตารางที่ 4.2: ผลการทดสอบประสิทธิภาพของการตรวจจับกรอบสีเหลี่ยมภายในวิดีโอ

จากตารางที่ 4.3 ผู้วิจัยได้ทำการทดสอบความแม่นยำและความเร็วในการประมวลผลของการใช้โมเดลปัญญาประดิษฐ์ YOLO-v3 320 ประมวลผลทุกเฟรม แม้จะตั้งขอบเขตอัตราส่วนร่วมของกรอบที่เหลืออยู่ที่จะนับว่าการทำนายถูกต้องสูงถึง 80% แต่ความแม่นยำสูงถึง 95% ใช้เวลาในการประมวลผล 452 วินาที เนื่องจากเฟรมละ 1.2 วินาที ซึ่งถือเป็นความแม่นยำที่สูงมากเมื่อเทียบกับเวลาที่ใช้ในการประมวลผลโดยใช้โมเดลปัญญาประดิษฐ์ประมวลผลเฉพาะบางเฟรมทุกๆ 1 วินาที แล้วใช้ระบบการทำนายตำแหน่งต่อไปของวัตถุในการสร้างกรอบสีเหลี่ยมในเฟรมระหว่างนั้น 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%

ต่อมาเป็นการทดสอบโดยใช้โมเดลปัญญาประดิษฐ์ประมวลผลเฉพาะบางเฟรมทุกๆ 1 วินาที แล้วใช้ระบบการทำนายตำแหน่งต่อไปของวัตถุในการสร้างกรอบสีเหลี่ยมในเฟรมระหว่างนั้น เพื่อเพิ่มความเร็วในการประมวลผล โดยระยะที่ใช้ในการทดสอบคือ ทุกๆ 10 เฟรม 20 เฟรม และ 25 เฟรม ซึ่งจากการทดสอบนั้นพบว่ามีการนี้มีความแม่นยำลดลงมาเหลือ 85% น้อยกว่าอยู่เพียง 10% เท่านั้น ถือเป็นความแม่นยำที่สูงเมื่อเทียบกันด้วยระยะเวลาในการประมวลผล ในขณะที่การใช้ระยะประมวลผลเป็น 10 เฟรมนั้นใช้เวลาในการประมวลผลเพียง 69 วินาที น้อยกว่าการใช้โมเดลปัญญาประดิษฐ์ YOLO-v3 320 ประมวลผลทุกเฟรมถึง 383 วินาที ซึ่งเร็วกว่าถึง 6.5 เท่า และความแม่นยำลดลงมาเหลือ 85% น้อยกว่าอยู่เพียง 10% เท่านั้น ถือเป็นความแม่นยำที่สูงเมื่อเทียบกันด้วยระยะเวลาในการประมวลผล ในขณะที่การใช้ระยะประมวลผล 20 เฟรมนั้นจะประมวลผลเร็วกว่าการใช้โมเดลปัญญาประดิษฐ์ YOLO-v3 320 ประมวลผลทุกเฟรมถึง 11 เท่า และมีความแม่นยำต่ำกว่า 15% และเมื่อใช้ระยะประมวลผล 25 เฟรมจะเร็วกว่าประมาณ 13 เท่า และความแม่นยำต่ำลงถึง 20%

4.4 ผลการทดสอบระบบระบุตัวตนของมนุษย์

4.4.1 ทดสอบประสิทธิภาพการทำงานของโมเดลปัญญาประดิษฐ์สำหรับการระบุตัวตนของบุคคล

ความแม่นยำของโมเดลปัญญาประดิษฐ์จากแหล่งที่มีมาค่าดังตารางด้านล่างนี้ ต่อมานำโมเดลปัญญา

| 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

ประดิษฐ์แต่ละอันมาทดสอบกับตัวอย่างภาพชุดข้อมูลที่ทางคณะผู้วิจัยได้สร้างขึ้น โดยภาพชุดข้อมูลที่นำมาใช้จะผ่านการตรวจหาบุคคลภายในภาพด้วยโมเดลปัญญาประดิษฐ์ YOLO v3 320



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 | 0.3 042 |
| 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 | 0.3 134 |
| ResNet50 MSMT17 | 0.3968 |

ตารางที่ 4.6: ผลการทดสอบความแม่นยำสำหรับการระบุบุคคลของโมเดลปัญญาประดิษฐ์ครั้งที่ 3

ค่าความแม่นยำในการระบุบุคคลนั้นค่าอยู่ที่ 0 แสดงบุคคลใน 2 เฟรมนั้นเป็นบุคคลเดียวกัน จากการทดลองครั้งที่ 1 จะเห็นเฟรมที่ไม่ต่อเนื่องกัน การทดลองครั้งที่ 2 และ 3 นั้นจะเป็นเฟรมที่ต่อเนื่องกันมากขึ้น ตามลำดับ ซึ่งจะแสดงให้เห็นว่าโมเดลปัญญาประดิษฐ์ที่สามารถให้ผลลัพธ์ที่มีประสิทธิภาพต่อเนื่องมากที่สุดคือ ResNet50 Market1501

4.5 ผลการทดสอบการจัดการกระทำของมนุษย์

4.5.1 ทดสอบประสิทธิภาพการทำงานของโมเดลปัญญาประดิษฐ์ที่เคยถูกเทรน์ผ่าน AVA เทียบผลลัพธ์กับแหล่งอ้างอิง ได้ผลการทดลองดังตารางต่อไปนี้

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

ตารางที่ 4.7: ผลการทดสอบความแม่นยำของโมเดลปัญญาประดิษฐ์เทียบผลลัพธ์กับแหล่งอ้างอิง

ความเร็วของต่อรูปภาพทางผู้วิจัยได้ใช้กราฟฟิกการ์ด GTX 2080 Ti ในการทดสอบซึ่งจะให้ความเร็วอยู่ที่ 5 วินาที ซึ่งทางแหล่งอ้างอิงนั้นใช้กราฟฟิกการ์ด Nvidia GeForce GTX TITAN X ในส่วนของค่าความแม่นยำที่ไม่เท่ากัน คาดว่าจะเป็นเพราะการประมวลผลของกราฟฟิกการ์ดของรุ่นที่ต่างกันจึงทำให้ค่า mAP ที่ออกมากไม่เท่ากัน

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

ตารางที่ 4.8: ผลการทดสอบความแม่นยำของโมเดลปัญญาประดิษฐ์ เมื่อใช้กับชุดข้อมูลที่ผู้วิจัยสร้างขึ้น

4.5.2 ผลการทดสอบประสิทธิภาพการทำงานของโมเดลปัญญาประดิษฐ์ที่เคยถูกเทรนด์ผ่าน AVA และ ใช้ชุดข้อมูลที่ผู้วิจัยสร้างขึ้น ใน การทดสอบและเทียบผลลัพธ์กับแหล่งอ้างอิง

4.5.3 ทดสอบประสิทธิภาพการทำงานของโมเดลปัญญาประดิษฐ์ที่เคยถูกเทรนด์ผ่านชุดข้อมูลสำหรับการเทรนด์ที่ผู้วิจัยสร้างขึ้น และ ใช้ชุดข้อมูลที่ผู้วิจัยสร้างขึ้น ใน การทดสอบและเทียบผลลัพธ์การทดสอบก่อนหน้า

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

ตารางที่ 4.9: ผลการทดสอบความแม่นยำของโมเดลปัญญาประดิษฐ์ที่ผู้วิจัยสร้างขึ้น ใช้กับชุดข้อมูลที่ผู้วิจัยสร้างขึ้น

ເອກສາຣອ້າງອີງ

- [1] Optical flow.
- [2] 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.
- [3] 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.
- [4] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset, 2018.
- [5] 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.
- [6] Gunnar Farnebäck. Two-frame motion estimation based on polynomial expansion. In *Scandinavian conference on Image analysis*, pages 363–370. Springer, 2003.
- [7] 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.
- [8] 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.
- [9] 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.
- [10] 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.
- [11] Karin Kipper Schuler. *Verbnet: A Broad-coverage, Comprehensive Verb Lexicon*. PhD thesis, Philadelphia, PA, USA, 2005. AAI3179808.

- [12] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions, 2014.
- [13] 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.
- [14] 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.
- [15] 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

ตัวอย่างชุดข้อมูลที่ผู้วิจัยสร้างขึ้น

ตัวอย่างชุดข้อมูลสำหรับการทดสอบโมเดลปัญญาประดิษฐ์ในการตรวจจับภาพบุคคล



รูปที่ ก.1: รูปผลลัพธ์การทำงานของหน้าต่าง Track