สรุปการทำงาน

ระบบนี้ออกแบบมาเพื่อวิเคราะห์พฤติกรรมของมนุษย์แบบหลายคนในวิดีโอแบบเรียลไทม์ โดยมีขั้นตอนการทำงานหลักดังนี้:

1. **ตรวจจับและติดตามบุคคล**: ใช้ YOLOv11 ร่วมกับ ByteTrack เพื่อระบุตำแหน่งและติดตามบุคคลในวิดีโอ พร้อมกำหนด track_id สำหรับแต่ละบุคคลในแต่ละเฟรม
2. **ประมวลผล keypoints ด้วย MediaPipe**: สำหรับ bounding box ที่ตรวจพบว่าเป็นมนุษย์ (label = "human") ระบบจะใช้ MediaPipe Pose ในการหา keypoints (joint coordinates) ของร่างกายในรูปแบบ skeleton
3. **ปรับ keypoints ให้อยู่ในรูปแบบ NTU RGB+D**: เนื่องจากชุดข้อมูล NTU RGB+D ใช้จุด joint ที่ต่างจาก MediaPipe จึงต้องทำการ mapping และ normalize keypoints เพื่อให้ตรงตามมาตรฐาน skeleton vector ของ NTU
4. **พยากรณ์พฤติกรรมด้วย LSTM**: ข้อมูล keypoints ที่ถูกเก็บต่อเนื่อง 30 เฟรม (1 sequence) จะถูกส่งเข้า LSTM Model ที่ผ่านการฝึกมาแล้ว เพื่อจำแนกพฤติกรรม เช่น ยืนอยู่ (standing), เคลื่อนไหว (moving), หรือกำลังยกของ (carrying)
5. **แสดงผลแบบเรียลไทม์**: แสดง Bounding Box, ID, Confidence Score และ Action Label บนวิดีโอสดด้วย OpenCV เพื่อให้เห็นพฤติกรรมของแต่ละบุคคลพร้อม ID

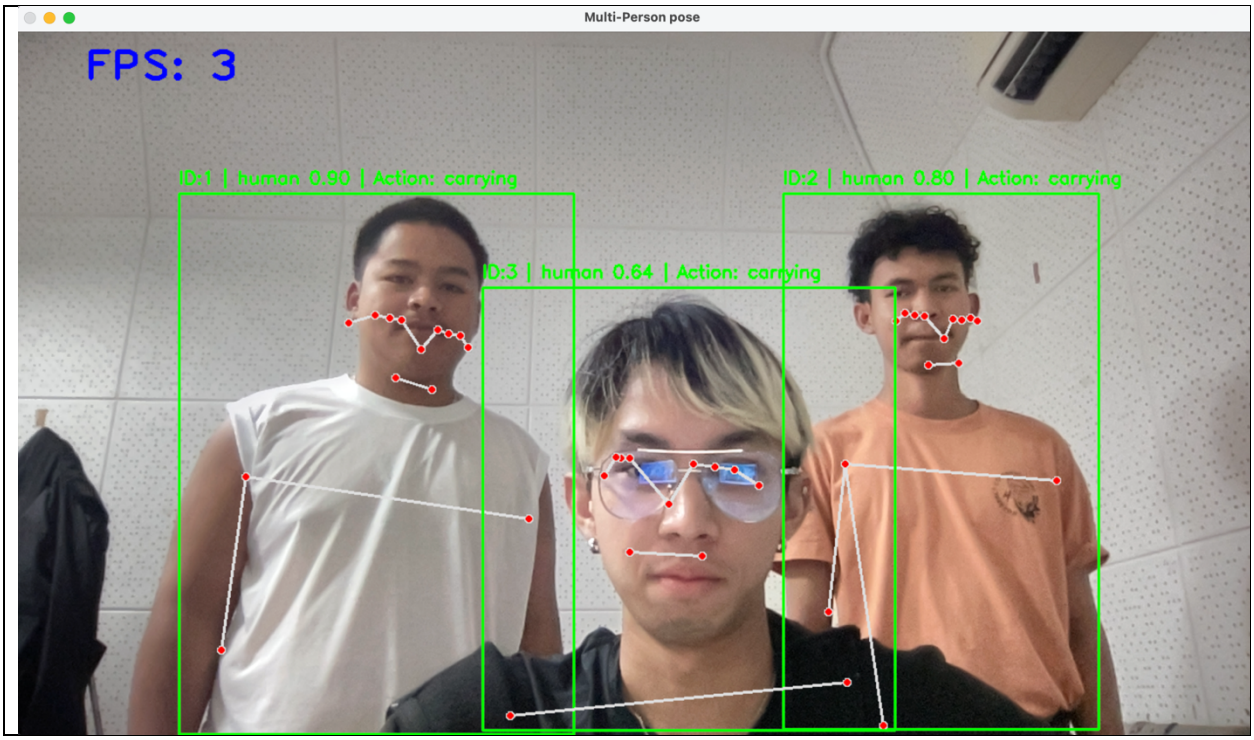
การเตรียมข้อมูลและการฝึกโมเดล

- ทำการ **data preprocessing** ข้อมูล .skeleton จาก NTU RGB+D dataset อย่างละเอียด โดยแปลงเป็น .npz สำหรับใช้กับ LSTM Model
- ทดลอง **เทรนโมเดล LSTM** ทั้งหมด 10 แบบ โดยปรับพารามิเตอร์ เช่น จำนวน nodes, จำนวน layers, learning rate เพื่อหา configuration ที่ให้ผลลัพธ์ดีที่สุด
- ทำการ **preprocessing ข้อมูลใหม่** ก่อนนำไปใช้จริง เพื่อให้แน่ใจว่าข้อมูลมีความสะอาดและตรงกับ input format ของโมเดล

รายละเอียดของระบบ

หัวข้อ	**รายละเอียด**
YOLO + ByteTrack	ตรวจจับบุคคลหลายคนพร้อมการติดตาม ID แต่ละคน
MediaPipe Pose	ดึงจุด skeleton แบบแม่นยำภายใน bounding box
Normalization	ปรับ keypoints ให้ตรงกับ NTU RGB+D เพื่อให้เข้าโมเดล LSTM ได้ถูกต้อง
Per-person Buffer	ใช้ buffers[track_id] ในการเก็บข้อมูลแยกตามบุคคล
Real-time Visualization	แสดง action + ID แบบ overlay ทันที

ตัวอย่างผลลัพธ์ในเบื้องต้น



สรุปผลลัพธ์เบื้องต้นของระบบ Multi-person Action Recognition

ภาพนี้แสดงตัวอย่างผลลัพธ์จากระบบที่ใช้ YOLOv11 + ByteTrack ร่วมกับ MediaPipe Pose และ LSTM Model โดยสามารถ:

- ตรวจจับบุคคลในภาพได้ครบทั้ง 3 คน พร้อมระบุ ID, confidence, และ action label
- แสดงผล action แบบ overlay พร้อมโครงร่าง skeleton keypoints บนร่างกายแต่ละบุคคลแบบ real-time

ข้อสังเกตจากผลลัพธ์

1. **การตรวจจับบุคคลและ tracking ID ทำงานได้ถูกต้อง**: ระบบสามารถแยกบุคคล 3 คนออกจากกัน และกำหนด ID ไม่ซ้ำกันได้แม่นยำ
2. **การพยากรณ์ Action ยังไม่แม่นยำ**:
    - ทุกคนถูกประเมินว่า "carrying" ทั้งที่ไม่มีใครถือของจริง ⟶ แสดงว่าโมเดลยังแยกพฤติกรรมได้ไม่ดีพอ
    - Confidence ของบางคนค่อนข้างต่ำ (เช่น ID:3 มี confidence = 0.64) ซึ่งอาจมาจาก keypoints ไม่ครบหรือมี noise
3. **ความละเอียดของ Pose ยังไม่เสถียร**:
    - Keypoints บางจุดไม่ตรงตามร่างกายจริง เช่น บริเวณแขนหรือขา ⟶ อาจเกิดจากการ crop ROI หรือ lighting

แนวทางปรับปรุง

- **เพิ่มความหลากหลายของข้อมูลฝึก (Training Data Augmentation)** เช่น ท่า postures, มุมกล้อง, การแต่งตัว
- **ปรับ label training ใหม่ให้สะท้อน action จริง** เช่น standing, idle, talking แยกจาก carrying
- **ใช้ Threshold ควบคุมการ predict**: ถ้า confidence ต่ำ ให้ label เป็น "unknown" แทน
- **ใส่ action log per ID เพื่อดูว่า action เปลี่ยนจริงหรือไม่ในช่วงเวลา**