Human Activity Recognition

By: Aditi Srivastava

Overview

Human Activity Recognition (HAR) involves identifying physical movements of humans such as walking, running, sitting, jumping, and waving. This project uses pose estimation and deep learning to classify actions in real-time using webcam input.

Problem Statement

The goal is to detect and classify human activities such as walking, jogging, running, hand-waving, boxing, and clapping from video frames. Traditional CNN approaches struggle because they analyze single images and fail to understand motion.

Activities:

- Walking
- Running
- Boxing
- Hand Waving, Jogging, Clapping

Why CNN Fails

CNN processes frames independently and cannot capture sequential motion. Even averaging predictions across frames produces inaccurate classification for actions requiring temporal understanding.

CNN Limitation:

Single-frame analysis → No motion understanding

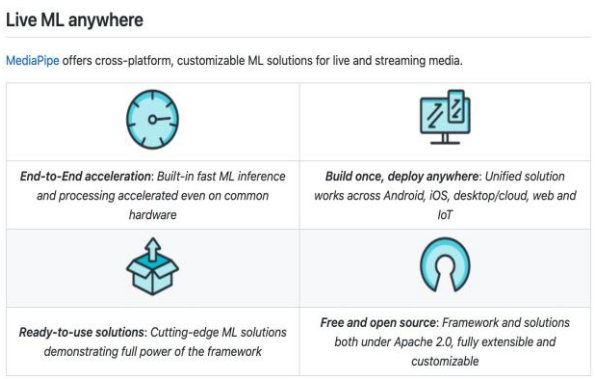
Training Pipeline

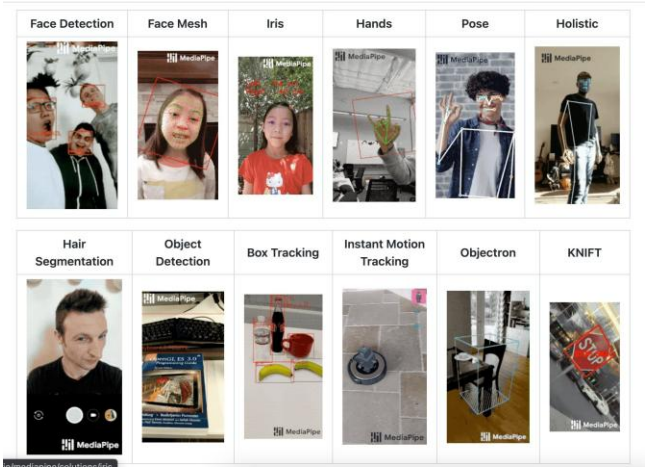
1. Capture video frames from webcam.
2. Extract body keypoints using MediaPipe Pose.
3. Store extracted keypoints into CSV files.
4. Build dataset containing sequences of poses.
5. Train an LSTM deep learning model. 6. Save the trained model for inference.

Training Pipeline Diagram
Video -> Pose Extraction -> CSV Dataset
CSV -> LSTM Training ->Saved Model

Real-time ready model

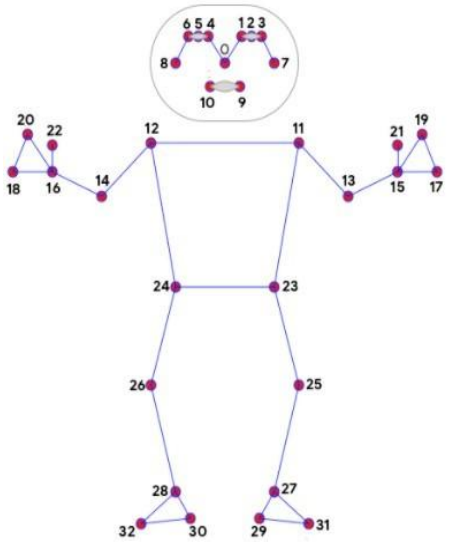
MediaPipe





MediaPipe Pose

MediaPipe Pose provides 33 body landmarks with real-time performance. It extracts skeleton keypoints that represent human posture and motion essential for understanding activities.



0. nose
1. left_eye_inner
2. left_eye
3. left_eye_outer
4. right_eye_inner
5. right_eye
6. right_eye_outer
7. left_ear
8. right_ear
9. mouth_left
10. mouth_right
11. left_shoulder
12. right_shoulder
13. left_elbow
14. right_elbow
15. left_wrist
16. right_wrist
17. left_pinky
18. right_pinky
19. left_index
20. right_index
21. left_thumb
22. right_thumb
23. left_hip
24. right_hip
25. left_knee
26. right_knee
27. left_ankle
28. right_ankle
29. left_heel
30. right_heel
31. left_foot_index
32. right_foot_index

Why LSTM?

LSTM (Long Short-Term Memory) networks are excellent for sequential data. Human activities rely on how movement changes across frames. LSTM learns motion patterns over time, making it ideal for HAR.

Why LSTM Works:

- Captures sequential motion
- Remembers long-term patterns
- More accurate than CNN for actions

Inference Pipeline

1. Read frame from webcam
2. Extract pose landmarks using MediaPipe
3. Structure keypoints into a sequence
4. Feed the sequence into the trained LSTM
5. Predict action and show label on screen

Inference Pipeline

Webcam Frame → Pose Extractor

Pose Landmarks → LSTM Model

Prediction → Display on Screen

Demo Output

The system displays the detected skeleton and predicted activity (e.g., Walking, Running, Jumping) in real time. This makes it suitable for monitoring, safety, fitness, and robotics applications.