UNIVERSIDAD DEL VALLE DE GUATEMALA
Visión por Computadora
Sección 10
Dr. Alan Gerardo Reyes

Proyecto #2

José Pablo Orellana - 21970
Diego Alberto Leiva – 21752
María Marta Ramírez - 21342
Gustavo González - 21438

Guatemala 10 de abril, 2025

Introducción

El presente proyecto se centra en la aplicación de técnicas de visión computacional en tiempo real, utilizando la biblioteca MediaPipe, desarrollada por Google. El objetivo principal es detectar y visualizar keypoints anatómicos humanos en distintos escenarios, lo cual es esencial para tareas como reconocimiento de gestos, seguimiento de movimientos, y análisis biomecánico.

Este trabajo se divide en dos partes complementarias:

- Parte 1: Detección de Manos y Keypoints (Hands)
  Se procesan imágenes estáticas para detectar manos y sus puntos clave anatómicos (landmarks). Se guardan las coordenadas 3D en archivos de texto y se genera una visualización con las conexiones entre los puntos.
- Parte 2: Estimación de Pose Corporal en Video
  Se analiza un video donde dos personas bailan, para estimar la pose corporal cuadro a cuadro. Se generan visualizaciones que muestran la superposición del esqueleto sobre el video original y una vista alternativa solo del esqueleto sobre fondo negro.

Ambas etapas demuestran la capacidad de usar modelos preentrenados de MediaPipe para obtener resultados en tiempo real, con buena precisión y bajo consumo de recursos computacionales, todo esto utilizando Python y herramientas de procesamiento de video e imagen.

Objetivo General

Desarrollar una aplicación capaz de detectar y visualizar puntos clave de manos y del cuerpo humano en imágenes y videos mediante el uso de herramientas como MediaPipe, OpenCV y NumPy.

Objetivos Específicos

- Detectar la presencia de manos humanas en imágenes estáticas y extraer 21 puntos clave por mano.
- Dibujar los landmarks y sus conexiones sobre las imágenes procesadas.
- Aplicar estimación de pose corporal a un video de entre 20 y 30 segundos de duración.
- Detectar los puntos clave del cuerpo humano y representar las conexiones como un esqueleto.
- Generar una visualización dual: una con el video original y esqueleto superpuesto, y otra con fondo negro mostrando únicamente el esqueleto.

Resultados

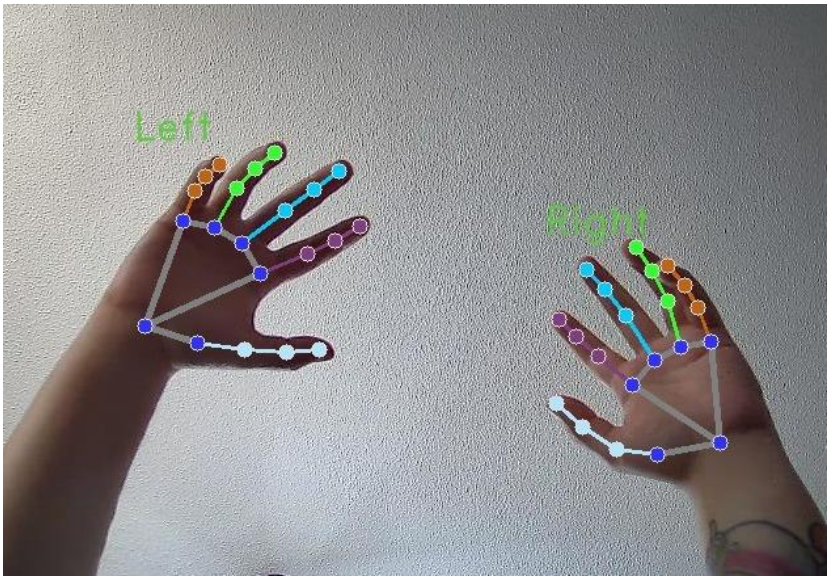

Imagen 1. Detección de mano derecha e izquierda.

Imagen 2. Detección de 21 keypoints de la mano.

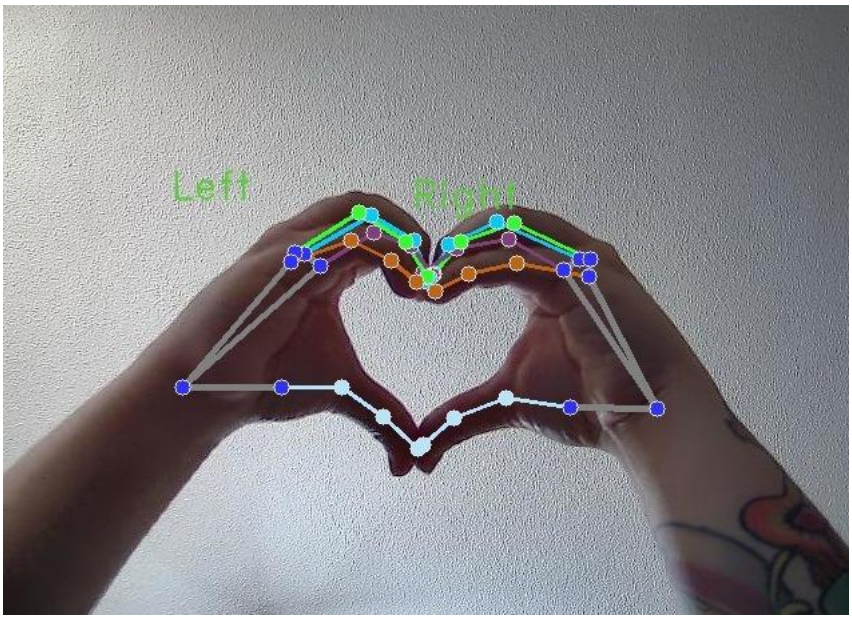

Imagen 3. Detección de figura con hand tracking.

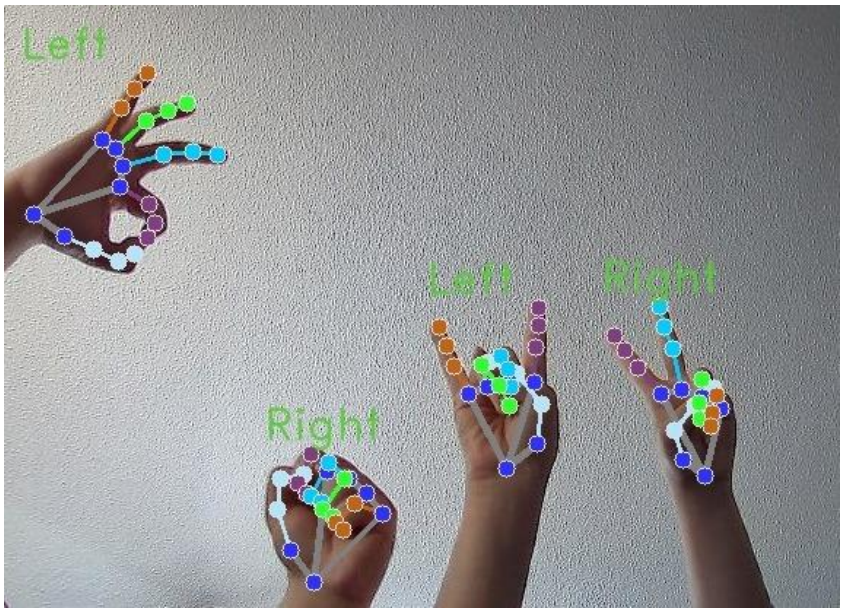

Imagen 4. Detección de varias manos en simultáneo.

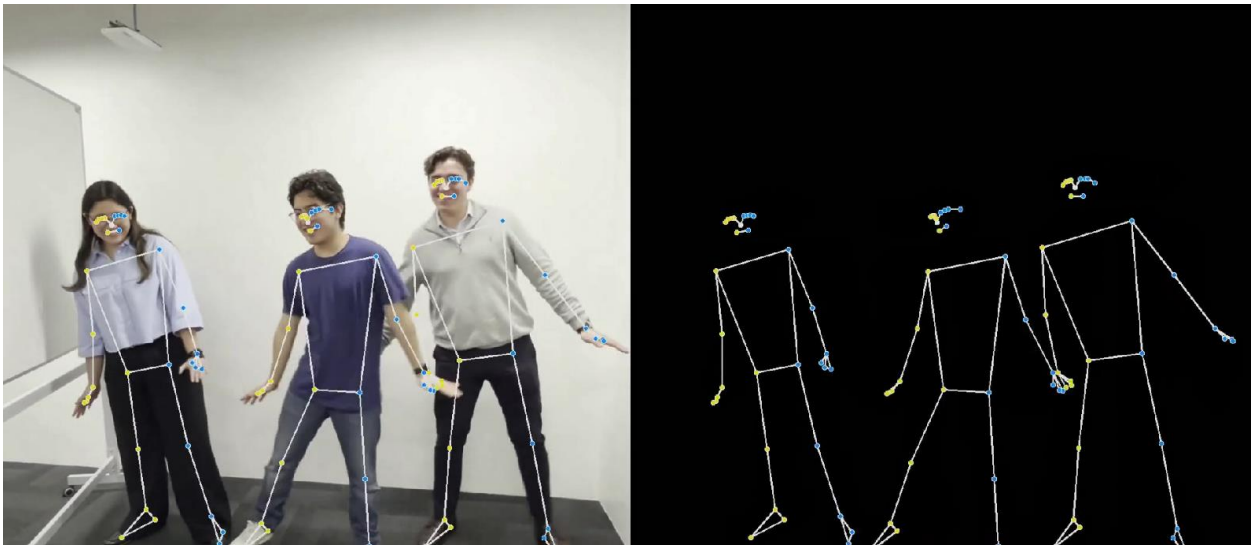

Imagen 5. Pose tracking con el video del grupo bailando.

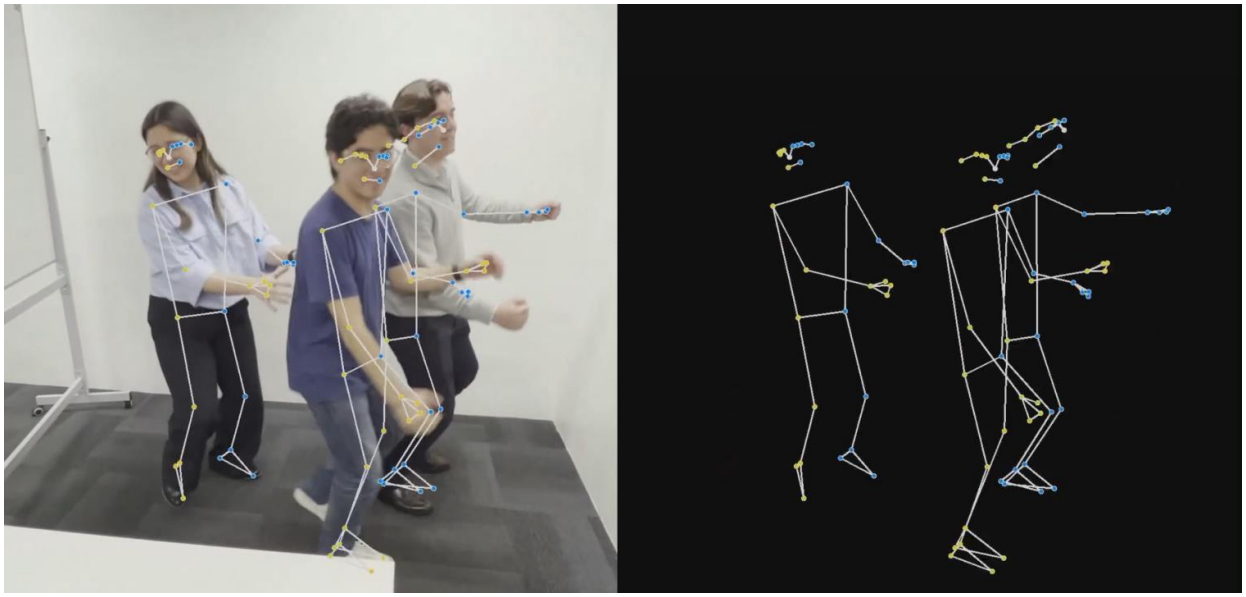

Imagen 6. Correcta visualización de la superposición de personas con el modelo.

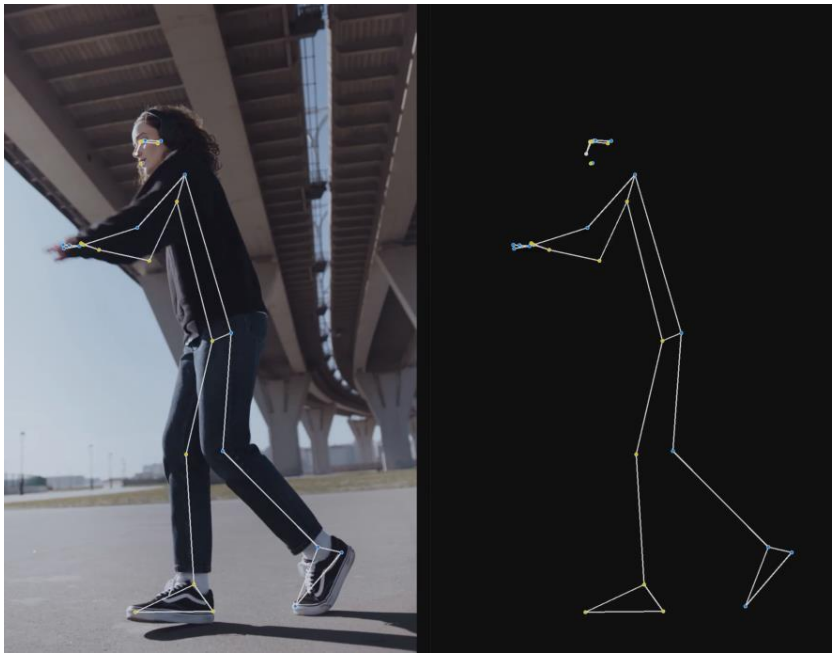

Imagen 3. Pose tracking con video de ejemplo

Discusión

En la detección de manos, los keypoints se identificaron con alta precisión incluso en imágenes con fondos complejos o manos en diversas orientaciones. No se presentaron errores significativos, y el rendimiento del modelo fue satisfactorio en múltiples condiciones. Para esta tarea se utilizaron los valores predeterminados de MediaPipe Tasks: una confianza mínima de 0.5 para la detección de la mano, presencia de la mano y rastreo (min_hand_detection_confidence, min_hand_presence_confidence, min_tracking_confidence). Estos valores fueron suficientes para obtener resultados precisos sin necesidad de ajustes adicionales.

El modo de ejecución elegido fue VIDEO, el cual permite un análisis secuencial del video completo. Aunque se probó inicialmente el modo LIVE_STREAM, se descartó por introducir un retardo notable en la transmisión, debido a que este modo opera mediante callbacks, lo que genera una sensación de lentitud similar a una cámara lenta. También se probó una versión legacy de MediaPipe (la solución anterior conocida como "MediaPipe Solutions"), que si bien era más liviana y fácil de integrar, tenía limitaciones significativas: únicamente permitía detectar un máximo de dos manos, y ofrecía menos parámetros configurables. Finalmente, se optó por la versión moderna (MediaPipe Tasks), que permite una detección más robusta y la posibilidad de rastrear hasta 10 manos, configuradas explícitamente para evitar sobrecargar el sistema.

En cuanto a la estimación de pose corporal en video, se obtuvieron resultados igualmente positivos. El modelo fue capaz de seguir con precisión los movimientos de los sujetos durante una coreografía de 20–30 segundos, manteniendo la continuidad entre cuadros y evitando saltos o

pérdidas de seguimiento. Para lograr este desempeño, se realizaron varios ajustes clave en la configuración del Pose Landmarker:

- Se incrementó la confianza mínima para detección de pose de 0.5 a 0.6 (min_pose_detection_confidence)
- Se incrementó la confianza mínima para presencia de pose de 0.5 a 0.7 (min_pose_presence_confidence)
- Se incrementó la confianza mínima para rastreo de 0.5 a 0.7 (min_tracking_confidence)
- Se estableció el modo de ejecución VIDEO en el parámetro VisionRunningMode, lo cual asegura un análisis cuadro por cuadro con coherencia temporal

Estos cambios permitieron una mayor precisión y estabilidad en la detección, especialmente en escenas donde los cuerpos de los participantes se cruzaban o interactuaban físicamente. Aunque inicialmente se observó que el modelo fallaba al detectar correctamente a los tres bailarines simultáneamente (solo mostraba uno o dos esqueletos en varios cuadros), los ajustes realizados mejoraron significativamente la capacidad del sistema para mantener un seguimiento consistente de múltiples personas.

La implementación también incluyó una vista dual del video: una versión con el esqueleto superpuesto sobre la imagen original y otra con un fondo negro que resalta únicamente los landmarks. Esta visualización permitió apreciar con mayor claridad el desempeño del modelo frente a variaciones de movimiento, orientación corporal y contacto entre los sujetos.

Los resultados validan la eficiencia de los modelos HandLandmarker y PoseLandmarker, y demuestran cómo una correcta configuración (apoyada en la documentación técnica oficial de MediaPipe) puede maximizar el rendimiento en aplicaciones del mundo real, especialmente en contextos dinámicos como el análisis de movimiento humano en video.

Conclusiones

- Se logró implementar la detección de manos y cuerpo utilizando modelos preentrenados de MediaPipe.
- La visualización gráfica de los landmarks permite analizar posturas y movimientos de forma clara.
- La integración con OpenCV facilitó el procesamiento de imágenes y videos en tiempo real.
- Se cumplieron los objetivos propuestos, demostrando la efectividad del enfoque basado en visión por computadora.

Referencias

- Google. (s.f.). MediaPipe Hand Landmarker. Google Developers. Recuperado el 10 de abril de 2025, de https://developers.google.com/mediapipe/solutions/vision/hand_landmarker
- Google. (s.f.). MediaPipe Pose Landmarker. Google Developers. Recuperado el 10 de abril de 2025, de https://developers.google.com/mediapipe/solutions/vision/pose_landmarker
- Google. (s.f.). MediaPipe Pose Landmarker - Configuration Options. Google AI. Recuperado el 10 de abril de 2025, de https://ai.google.dev/edge/mediapipe/solutions/vision/pose_landmarker?hl=es-419#configurations_options
- OpenCV. (s.f.). OpenCV Documentation. OpenCV.org. Recuperado el 10 de abril de 2025, de https://docs.opencv.org
- NumPy. (s.f.). NumPy Documentation. NumPy.org. Recuperado el 10 de abril de 2025, de https://numpy.org/doc/