Sistema de Clasificación de Actividades Humanas Mediante Análisis de Landmarks y Modelos de Machine Learning

Cristian Molina
Carlos Sanchez
Juan Esteban Eraso
Damy Villegas

Universidad ICESI - APO3 - 2025 -2

Abstract: **This project presents the development of an intelligent system capable of analyzing human activities from video sequences through the extraction of anatomical landmarks using MediaPipe and the subsequent classification of these movements with machine learning techniques. The system computes biomechanical features—including joint angles, limb velocities, lateral trunk inclination, and frame-to-frame motion deltas—to enhance the discriminative power of the dataset. An adapted CRISP-DM methodology was applied, covering data collection, preprocessing, feature engineering, model selection, hyperparameter optimization, and performance evaluation using metrics such as accuracy, recall, and F1-score. The results demonstrate high classification performance across activities such as walking, turning, sitting, standing, and approaching. Additionally, the report includes a mathematical formulation of the computed metrics, an ethical analysis of the system's implications, and a discussion on potential global impacts of artificial-intelligence-based motion analysis.**

Resumen: ***Este proyecto presenta el desarrollo de un sistema capaz de analizar actividades humanas en tiempo real a partir de videos, mediante el seguimiento de articulaciones corporales usando MediaPipe y la posterior clasificación mediante algoritmos de machine learning. Se procesaron múltiples videos capturados por los integrantes del grupo, extrayendo landmarks (caderas, rodillas, tobillos, hombros, cabeza y muñecas), generando métricas biomecánicas como ángulos, velocidades articulares e inclinaciones laterales. Posteriormente, se implementó una metodología CRISP-DM adaptada que incluyó recolección de datos, preprocesamiento, ingeniería de características, selección de modelos, ajuste de hiper parámetros y evaluación mediante métricas como accuracy, recall y F1-score. Los resultados muestran altos niveles de precisión en la clasificación de actividades como caminar, girar, sentarse, levantarse y acercarse. Además, se presenta un análisis ético, matemático y de impacto global asociado con el uso de sistemas de visión artificial.***

I. INTRODUCCIÓN

El reconocimiento de actividades humanas es un problema relevante en campos como la rehabilitación, el análisis deportivo, la vigilancia, la asistencia a personas mayores y la interacción humano-computador. El caso de estudio propuesto por la asignatura APO 3 consiste en construir un sistema capaz de analizar en tiempo real actividades como caminar hacia la cámara, retroceder, girar, sentarse y ponerse de pie.

El objetivo del proyecto es desarrollar un sistema completo que capture video, procese los landmarks corporales, derive características biomecánicas relevantes y clasifique automáticamente la actividad realizada. Este problema implica retos como el ruido en la captura, la variabilidad de poses humanas, la normalización espacial y la necesidad de métricas adecuadas para evitar sobreajuste.

Este informe reúne todo el proceso seguido, desde el planteamiento del problema hasta el entrenamiento del modelo final, incluyendo un análisis de desempeño, impactos éticos, formulaciones matemáticas y propuestas de mejora.

II. MARCO TEÓRICO

A. Seguimiento de Articulaciones (Pose Stimation):

MediaPipe Pose proporciona 33 puntos clave del cuerpo humano. Para este proyecto se emplearon: cadera, rodillas, tobillos, hombros, cabeza y muñecas, permitiendo identificar posturas globales y movimientos dinámicos.

Cada landmark contiene coordenadas *(x, y, z)* normalizadas espacialmente respecto al frame en el que fueron capturadas.

B. Normalización Espacial:

La posición absoluta en la imagen depende de:

- Distancia del usuario a la cámara
- Altura del sujeto
- Ángulo de grabación
- Perspectiva

Para obtener métricas más robustas se normalizaron todos los landmarks respecto al centro de las caderas:

$$L_{norm} = \frac{L - C_{hip}}{\|C_{hip}\|}$$

Esto elimina dependencias externas y permite que el modelo aprenda únicamente el movimiento relativo.

C. Métricas Biomecánicas Calculadas:

Del conjunto de landmarks se calcularon:

1. Ángulos articulares

Por ejemplo, ángulo de la rodilla:

$$\theta = \arccos\left(\frac{(A - B)\cdot(C - B)}{\|A - B\|\|C - B\|}\right)$$

2. Velocidades articulares

$$v = \frac{p_t - p_{t-1}}{\Delta t}$$

3. Inclinación lateral del tronco

$$incl = |x_{shoulders} - x_{hips}|$$

4. Deltas de movimiento

Diferencia entre frames consecutivos para capturar dinámicas temporales.

D. Modelos de Machine Learning

Se evaluaron varios modelos:

- **Random Forest**
- **XGBoost**

El modelo Random Forest optimizado resultó ser el más robusto, gracias a su uso de *class_weight* balanceado y características biomecánicas refinadas, logrando alta estabilidad y generalización sin caer en overfitting.

III. METODOLOGÍA

El desarrollo del sistema siguió una metodología basada en **CRISP-DM**, adaptada para permitir iteración continua y refinamiento progresivo. El proyecto se organizó en **tres fases principales**, alineadas con las entregas del curso: recolección y anotación de datos, ingeniería de características y entrenamiento/optimización de modelos. Finalmente, se integró un sistema funcional para inferencia en tiempo real.

A. Data Collection and Annotation:

Para el entrenamiento del sistema se recolectaron **18 videos cortos** (duración entre 10–20 segundos, 29–60 FPS, resoluciones 480×848 y 464×832) en los que diferentes personas realizaron actividades básicas:

- Caminar hacia el frente
- Caminar hacia atrás
- Giro 180°
- Sentarse
- Ponerse de pie

Los videos se capturaron en entornos controlados con iluminación uniforme para reducir el ruido visual.

- **Anotación temporal de actividades**

La anotación se realizó en dos etapas:

I. **Plantillas automáticas iniciales** con 4 clases amplias:
sentado, de_pie, caminando, transición.

II. **Refinamiento manual con Label Studio**
Se ajustaron los segmentos para obtener 5 clases finales:
caminar adelante, caminar atrás, giro 180°, ponerse de pie, sentarse.

Cada actividad se representó mediante su intervalo temporal **[inicio, fin] en frames**, posteriormente convertido a segundos según el FPS de cada video. Esto permite sincronizar las etiquetas con los landmarks y generar una base de datos temporalmente consistente.

B. Preprocessing & Feature Engineering:

1. Extracción de landmarks

Para cada video anotado se extrajeron los landmarks corporales utilizando **MediaPipe Pose**, muestreando a **15 FPS** para mejorar la eficiencia.

Los landmarks incluyen coordenadas (x, y, z) para articulaciones como:

- Cabeza
- Hombros
- Caderas
- Rodillas
- Tobillos
- Muñecas

2. Normalización espacial

Se aplicó una normalización dividiendo cada coordenada por la **distancia euclidiana entre los hombros** (izquierdo y derecho). Esto garantiza invariancia ante:

- Altura de la persona
- Distancia a la cámara
- Variaciones de perspectiva

3. Generación de Características

A partir de las secuencias temporales normalizadas se diseñaron **16 características biomecánicas**, incluyendo:

- ***mean_speed, std_speed:*** velocidad global del cuerpo
- ***trunk_angle_deg, trunk_angle_var:*** orientación y estabilidad del tronco
- ***head_y_mean, head_y_min, head_y_range:*** amplitud vertical de cabeza
- ***hip_y_mean:*** nivel vertical promedio de cadera
- ***hip_x_disp, hip_x_path:*** desplazamiento y distancia horizontal
- ***avg_vertical_speed_hip, avg_horizontal_speed_hip:*** velocidades direccionales
- ***normalized_head_hip_distance_mean:*** estabilidad del eje cuerpo–cabeza
- ***seg_len:*** duración total del segmento

Estas características fueron diseñadas para capturar patrones biomecánicos robustos como postura, simetría, velocidad y trayectoria, complementando el análisis teórico previo.

C. Model Selection and Optimization:

La selección y optimización de modelos se realizó de forma **iterativa**, evaluando en cada etapa métricas como F1-macro y matriz de confusión.

I. Iteración 1 — Random Forest básico:

- Se entrenó un modelo Random Forest con las **3 características más importantes** según análisis preliminar:
mean_speed, std_speed, trunk_angle_deg.
- Resultado: **F1-macro = 0.69**.
- Conclusión: El modelo capturaba algunos patrones, pero era insuficiente.

II. Iteración 2 — Características avanzadas:

Se añadieron nuevas características como:

- Aceleración
- Ángulos articulares dinámicos
- Métricas de simetría lateral

Resultado: F1-macro = 0.58.
Conclusión: El aumento de complejidad introdujo ruido y redujo el desempeño.

III. Iteración 3 — SMOTE + XGBoost:

- Se aplicó **SMOTE** para balancear el dataset.
- Se probó **XGBoost** con validación cruzada.

Resultado: F1-macro ≈ 0.64, con confusiones notables entre clases similares.

Conclusión: El modelo era estable pero no capturaba bien las diferencias sutiles.

IV. Iteración Final — Random Forest optimizado

Se seleccionaron las **16 características refinadas** y se ajustó un Random Forest con:

- `n_estimators = 200`
- `max_depth = 12`
- `min_samples_split = 2`
- `max_features = 'sqrt'`
- `class_weight = 'balanced_subsample'`

Optimización realizada con **GridSearchCV** (5 folds estratificados).

Resultado Final:

F1-macro=0.71±0.05

Este modelo mostró el mejor equilibrio entre estabilidad, interpretabilidad y capacidad de generalización.

D. Real-Time Implementation:

El modelo final se integró en una aplicación Python capaz de procesar video en tiempo real:

- Captura de webcam a 15 FPS
- Extracción de landmarks y normalización
- Cálculo dinámico de características sobre una ventana deslizante de 15 frames
- Predicción continua de la actividad
- Interfaz con:
    - Video en vivo
    - Esqueleto superpuesto
    - Actividad actual
    - Nivel de confianza

Se implementó suavizado temporal (temporal smoothing) para evitar predicciones erráticas entre frames consecutivos.

IV. RESULTADOS

A. Modelo Basado en Posiciones Absolutas:

La primera aproximación consistió en entrenar modelos supervisados usando exclusivamente las coordenadas absolutas (x, y, z) obtenidas directamente de MediaPipe Pose para cada articulación relevante. Este enfoque presenta varias limitaciones inherentes:

1. Dependencia del punto de vista de la cámara
   Las posiciones absolutas varían fuertemente según:
    - La distancia entre la persona y la cámara,
    - La altura del sujeto,
    - El ángulo de grabación,
    - La iluminación,
    - La calidad del video.
2. Variabilidad no relacionada con el movimiento
   Los modelos aprendían patrones irrelevantes (por ejemplo: *"persona aparece más grande → está cerca → probablemente caminando hacia el frente"*).
3. Sobreajuste severo
   A pesar de alcanzar altos puntajes en entrenamiento y validación interna, el modelo fallaba en generalizar a videos nuevos.

B. Resultados Cuantitativos:

El mejor modelo basado en posiciones fue un Random Forest entrenado con todas las coordenadas normalizadas por frame:

- Accuracy entrenamiento: > 95%
- Accuracy validación: ~ 98%
- Pruebas reales: desempeño deficiente, predicciones ruidosas
- Problema dominante: overfitting

Aunque los números eran altos, el modelo no capturaba la esencia biomecánica del movimiento y dependía excesivamente del contexto visual.

C. Conclusión del modelo basado en posiciones:

Los resultados demostraron que los valores absolutos de las articulaciones no son adecuados para tareas de reconocimiento de actividades. El modelo era sensible a cualquier variación externa y no servía en un entorno real.

Este análisis motivó la transición a un enfoque centrado en características derivadas que

representarán la biomecánica del movimiento en lugar de las posiciones pixeladas.

D. Visualización:

El sistema muestra:

- Actividad detectada en tiempo real.
- Ángulos de rodilla/cadera.
- Inclinación lateral.

E. Análisis de Resultados:

F. Análisis Ético y de Impacto:

I. Situaciones éticas:

1. Privacidad de datos de video.
2. Consenso informado.
3. Sesgos en personas grabadas.
4. Posible mal uso en vigilancia excesiva.

II. Relación con el Código de Ética:

1. Privacidad: Respeto por la información personal
2. Sesgos: Justicia y equidad
3. Seguridad: Bienestar público
4. Uso del sistema: Responsabilidad profesional

III. Impactos Globales:

Se elaboró una matriz como solicita la competencia PI2:

1) Social
   i. Impacto positivo: Rehabilitación, asistencia.
   ii. Riesgo o problema: Vigilancia.

2) Económico
   i. Impacto positivo: Bajo costo, aplicable comercialmente.
   ii. Riesgo o problema: Requerimiento de hardware.

3) Ambiental
   i. Riesgo o problema: Grabación continua.
   ii. Impacto positivo: Bajo consumo energético

4) Global
   iii. Impacto positivo: Transferible a otros países.
   iv. Riesgo o problema: Diferencias regulatorias.

G. Resolución de Problemas Matemáticos:

I. Formulación del problema

Dado un conjunto de secuencias de landmarks:

$$X = \{L_1, L_2, ..., L_n\}$$

predecir la actividad realizada:

$$f(X) \rightarrow y$$

donde $y \in$ {walk, turn, sit, stand, …}.

II. Justificación matemática

- Normalización vectorial para la invariancia espacial.
- Distancias euclidianas para medir movimiento.
- Cálculo de ángulos mediante producto punto.
- Modelos supervisados basados en optimización de funciones de pérdida:

$$L = \sum (y - \hat{y})^2$$

V. CONCLUSIONES

A. Logros:

1) Desarrollo de un sistema robusto basado en biomecánica.

   El uso de landmarks combinados con características derivadas (ángulos, velocidades, inclinación del tronco, desplazamientos) permitió construir un modelo mucho más estable y generalizable que las aproximaciones basadas únicamente en posiciones absolutas.

2) Optimización exitosa del modelo Random Forest.

   Gracias al refinamiento de 16

características y el uso de *class_weight balanced*, el Random Forest logró el mejor equilibrio entre precisión, estabilidad y capacidad de generalización (F1-macro ≈ 0.71 ± 0.05).

3) Metodología CRISP-DM aplicada correctamente.

La estructura iterativa permitió avanzar ordenadamente desde la recolección de datos hasta la implementación en tiempo real, refinando decisiones en cada etapa.

4) Normalización espacial efectiva.

La transformación de coordenadas permite eliminar variabilidad causada por distancia, altura y perspectiva, focalizando el aprendizaje en relaciones biomecánicas reales.

5) Sistema funcional en tiempo real.

La aplicación final integra webcam, esqueleto superpuesto y predicción continua con suavizado temporal, validando que el sistema es viable fuera del entorno experimental.

B. Trabajo Futuro:

1) Refinamiento de la Ingeniería de Características: Aunque el modelo logró buenos resultados generales, algunas actividades —especialmente *caminar hacia el frente*— siguen siendo difíciles de distinguir. En futuras versiones se propone diseñar características biomecánicas más específicas para los patrones de marcha, incorporando información más rica sobre la cadencia, la simetría del movimiento y la dinámica de las extremidades inferiores. Esto permitiría mejorar la capacidad del modelo para diferenciar actividades con trayectorias similares.

2) Incorporación de Modelos Temporales: Las actividades humanas tienen una naturaleza secuencial que no siempre es capturada completamente por modelos basados en características estáticas. Una línea de mejora consiste en evaluar modelos diseñados para manejar series temporales, como LSTM, GRU o Transformers. Estos enfoques podrían aprovechar mejor la continuidad del movimiento y mejorar el reconocimiento de actividades como caminar o girar.

3) Optimización de la Implementación en Tiempo Real: La integración en tiempo real mostró buenos resultados, pero aún puede optimizarse. Futuras versiones podrían incluir mecanismos de calibración automática, mejores visualizaciones dinámicas y una interfaz más interactiva orientada a aplicaciones clínicas, deportivas o de asistencia.

VI. REFERENCIAS

[I] Google, "*MediaPipe: Cross-platform, customizable ML solutions for live and streaming media,"* Google AI. [Online]. Available: https://developers.google.com/mediapipe

[II] G. Biau and E. Scornet, *"A random forest guided tour," Test*, vol. 25, no. 2, pp. 197–227, Apr. 2016. [Online]. Available: https://doi.org/10.1007/s11749-016-0481-7

[III] L. Breiman, *"Random Forests,"* Machine Learning, vol. 45, no. 1, pp. 5–32, 2001. [Online]. Available: https://doi.org/10.1023/A:1010933404324

[IV] T. Chen and C. Guestrin, *"XGBoost: A Scalable Tree Boosting System,"* in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785–794, 2016. [Online]. Available: https://doi.org/10.1145/2939672.2939785