Sistema de detección de actividades humanas en tiempo real mediante MediaPipe Pose y modelos de clasificación supervisada
Proyecto Final

Juan Camilo Molina Mussen – A00399775

Nicolas Cardona - A00373470

Daniel Gonzalez Rivera– A00399873

Sharik Camila Rueda Lucero – A00399189

Algoritmos y programación III

Ingeniería de Sistemas

Faculta Barberí de Ingeniería, Diseño y Ciencias Aplicadas

22 de noviembre de 2025

Santiago de Cali, Valle del Cauca, Colombia

Sistema de detección de actividades humanas en tiempo real mediante MediaPipe Pose y modelos de clasificación supervisada

Resumen

En este proyecto se desarrolla un sistema de visión por computador capaz de reconocer actividades humanas básicas a partir de secuencias de video y de analizar métricas posturales en tiempo real. El sistema utiliza MediaPipe Pose para extraer landmarks corporales de videos capturados por cámara, y a partir de estas posiciones articulares se construyen características agregadas como velocidad de la cadera, inclinación de los hombros, ángulo promedio de rodillas, movimiento inter-frame y brillo promedio. Para entrenar los modelos se conformó un conjunto de datos colaborativo, obtenido mediante el intercambio de videos entre grupos del curso, con distintas personas, entornos, condiciones de iluminación y velocidades de movimiento. Cada video se procesó de forma homogénea, generando un resumen numérico por acción etiquetada.

Sobre este conjunto de datos se evaluaron modelos supervisados clásicos (Random Forest, SVM y XGBoost) para clasificar cinco actividades: caminar hacia adelante, caminar hacia atrás, girar a la derecha, sentarse y pararse. El mejor desempeño en validación se obtuvo con Random Forest, con un F1 ponderado cercano a 0.83, seguido muy de cerca por SVM. Sin embargo, en el escenario de inferencia en tiempo real, integrado en una aplicación web desarrollada con Streamlit, se observó que el sistema predice de manera consistente las clases de caminar, pero presenta dificultades para distinguir de forma robusta las transiciones de sentarse y pararse. Este resultado evidencia que, aunque el enfoque propuesto es viable y funcional para ciertos tipos de movimiento, aún requiere mejorar la modelación temporal y la representación de las acciones más sutiles para lograr una generalización adecuada en condiciones reales de uso.

Introducción

El análisis automático del movimiento humano ha adquirido una importancia creciente en los últimos años debido a su aplicación en campos como la biomecánica, la rehabilitación física, la ergonomía laboral, la seguridad industrial y el deporte. Tradicionalmente, esta tarea ha requerido sistemas costosos de captura de movimiento, sensores corporales o equipos especializados difíciles de implementar en contextos cotidianos. El desarrollo reciente de herramientas de visión por computador basadas en aprendizaje automático ha permitido construir soluciones más accesibles que utilizan únicamente una cámara convencional y modelos de estimación de pose para extraer información relevante del cuerpo humano.

Para ello se emplea MediaPipe Pose, una solución de estimación de pose eficiente y de bajo costo computacional que permite extraer 33 puntos articulares del cuerpo con estabilidad suficiente para aplicaciones interactivas. A partir de estos landmarks se generan características cinemáticas y posturales que sirven como entrada para modelos supervisados de clasificación.

El problema que se aborda consiste en determinar, de forma automática, cuál de las actividades definidas está realizando una persona frente a una cámara, y en capturar variaciones relevantes en su postura corporal mediante el cálculo de métricas derivadas (velocidad de la cadera, inclinación del tronco, ángulos

articulares, entre otras). Este problema es desafiante debido a la variabilidad natural en la forma de moverse entre diferentes individuos, a los cambios en iluminación y perspectiva, y a la similitud visual entre ciertas actividades, especialmente las transiciones entre sentarse y pararse.

Teoria

El desarrollo del sistema propuesto se fundamenta en tres conceptos teóricos principales: la estimación de pose humana, la extracción de características cinemáticas a partir de landmarks y la clasificación supervisada de secuencias de movimiento. A continuación, se presentan los elementos teóricos esenciales para comprender el funcionamiento del proyecto.

1. Estimación de pose humana mediante MediaPipe Pose

MediaPipe Pose es un modelo de estimación de pose monocular que identifica 33 puntos articulares del cuerpo humano a partir de una sola imagen o frame de video. Cada landmark está definido por sus coordenadas normalizadas (x,y)(x,y) y un valor de visibilidad asociado. El modelo utiliza arquitecturas ligeras basadas en redes neuronales convolucionales para detectar los puntos corporales en tiempo real, incluso en dispositivos de bajo rendimiento.

Los landmarks relevantes para este proyecto incluyen hombros, caderas, rodillas, tobillos y cabeza, los cuales permiten describir la postura general y los movimientos principales del cuerpo. La estabilidad y consistencia de los landmarks convierten a MediaPipe en una alternativa eficiente para la captura de movimiento sin sensores físicos ni cámaras especiales.

2. Cálculo de características a partir de landmarks

A partir de los landmarks detectados se pueden derivar características (features) que representan propiedades cinemáticas y posturales del movimiento humano. Este proyecto emplea métricas diseñadas específicamente para capturar patrones entre actividades similares:

- Velocidad del centro de cadera: medida mediante la distancia euclidiana entre el centro de las caderas en frames consecutivos. Esta característica es útil para identificar actividades de desplazamiento como caminar.
- Inclinación de los hombros: diferencia vertical entre hombro izquierdo y derecho. Ayuda a detectar giros o posturas inclinadas del tronco.
- Ángulo promedio de rodillas: calculado mediante la geometría del triángulo cadera–rodilla–tobillo. Permite distinguir flexiones, como sentarse o prepararse para levantarse.
- Movimiento inter–frame de landmarks: promedio de desplazamientos de varios puntos clave del cuerpo entre frames consecutivos. Resume la intensidad del movimiento general.
- Brillo promedio del frame: utilizado como indicador simple de condiciones de iluminación, ya que éstas afectan la estabilidad de la estimación de pose.

Estas características se agregan sobre una ventana temporal de tamaño fijo, lo que permite suavizar el ruido de frame a frame y capturar tendencias más estables del movimiento, necesarias para distinguir entre acciones transicionales.

3. Clasificación supervisada de actividades humanas

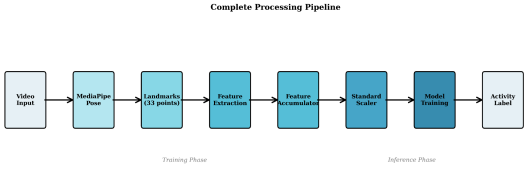
La tarea de clasificación consiste en asignar a cada ventana de características una etiqueta correspondiente a una actividad específica. Se emplean modelos supervisados, como Random Forest, Support Vector Machines y XGBoost, debido a su capacidad para trabajar con vectores de características tabulares y su efectividad en datasets relativamente pequeños.

- Random Forest utiliza múltiples árboles de decisión entrenados sobre submuestras del dataset, lo que lo hace robusto al ruido y al sobreajuste.
- SVM busca una frontera óptima entre clases en un espacio de alta dimensionalidad, siendo particularmente efectiva cuando las clases son separables mediante kernels no lineales.
- XGBoost implementa boosting de árboles de decisión, combinando modelos débiles para construir predictores más potentes.

En este proyecto, la representación tabular basada en promedios y variaciones de landmarks permite que estos algoritmos aprendan patrones distintivos asociados a cada actividad humana.

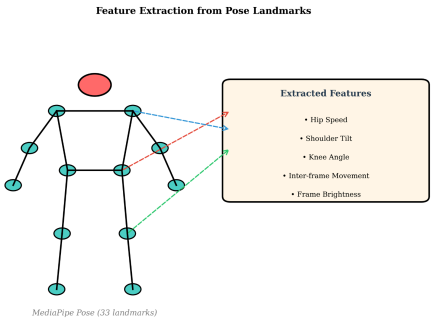
Metodología

El desarrollo del proyecto siguió un enfoque estructurado en cinco etapas principales.



Primero, se construyó el conjunto de datos mediante colaboración inter-grupal, donde cada equipo capturó videos de personas ejecutando las cinco actividades objetivo bajo condiciones variables de iluminación, fondo y distancia a cámara. Cada archivo fue etiquetado mediante un esquema uniforme basado en el nombre del archivo, lo que permitió automatizar la asignación de clases y garantizar consistencia entre los distintos orígenes de datos.

La segunda etapa consistió en emplear MediaPipe Pose para extraer 33 landmarks corporales por frame con coordenadas normalizadas y valores de confianza, herramienta seleccionada por su equilibrio entre precisión y eficiencia computacional, la extracción de características cinemáticas y posturales. El procesamiento operó frame por frame, convirtiendo coordenadas normalizadas a píxeles absolutos y filtrando puntos con confianza menor a 0.5. Se calcularon cinco métricas agregadas por video: velocidad promedio de cadera (distancia euclidiana entre frames consecutivos), inclinación promedio de hombros (diferencia vertical entre hombros), ángulo promedio de rodillas (cadera-rodilla-tobillo mediante productos vectoriales), movimiento promedio inter-frame (desplazamiento de articulaciones estables), y brillo promedio del frame.

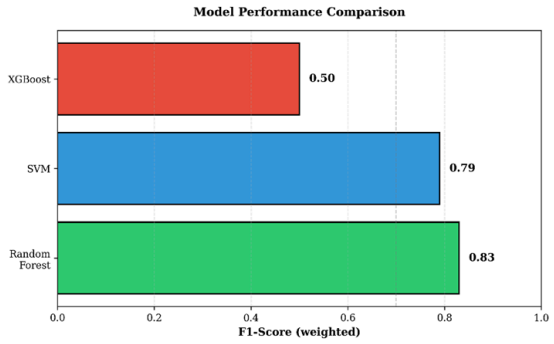


También se registraron el número de frames y duración en segundos. Esta

representación agregada simplifica la clasificación pero pierde información sobre dinámica temporal interna.

En la tercera etapa se realizó preprocesamiento y análisis exploratorio. Se eliminaron columnas no predictivas y se verificó ausencia de nulos y duplicados. El análisis reveló desbalance moderado entre clases y outliers en métricas de velocidad y ángulo de rodilla, los cuales se mantuvieron para preservar variabilidad natural. Se aplicó StandardScaler para normalizar variables numéricas y LabelEncoder para codificar las cinco clases. El dataset fue dividido en 80% entrenamiento y 20% prueba con estratificación por clase.

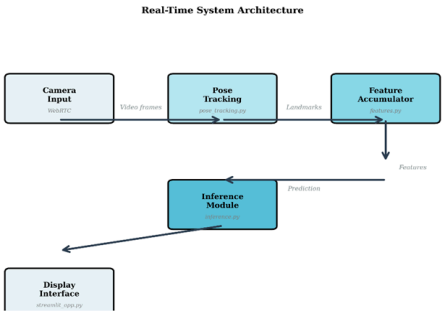
La cuarta etapa correspondió al entrenamiento de modelos supervisados. Se evaluaron Random Forest, SVM y XGBoost mediante GridSearchCV con validación cruzada de 3 folds, optimizando F1-score ponderado. Se exploraron hiperparámetros clave: en Random Forest (n_estimators: 100-300, max_depth: 5-20), en SVM (C: 0.1-10, kernel: RBF/poly), y en XGBoost (n_estimators: 100-200, learning_rate: 0.01-0.2). Se configuró class_weight='balanced' en Random Forest y SVM para mitigar desbalance. Los resultados mostraron que Random Forest alcanzó F1=0.83, SVM F1=0.79, y XGBoost F1=0.50.



Las matrices de confusión revelaron que ambos modelos principales clasificaban correctamente actividades de desplazamiento (caminar, girar), pero presentaban confusiones entre sentarse y pararse, sugiriendo que las características agregadas no capturan adecuadamente transiciones posturales breves.

Se seleccionó Random Forest como modelo final y fue serializado con LabelEncoder y StandardScaler mediante joblib.

La quinta etapa implementó inferencia en tiempo real mediante arquitectura modular de cuatro componentes:



(1) detección de pose con MediaPipe procesando frames a RGB y extrayendo landmarks, (2) acumulación de características mediante ventana deslizante de 15 frames calculando promedios móviles de las cinco métricas con mínimo de 10 frames para estabilidad, (3) predicción normalizando el vector con StandardScaler y aplicando votación por mayoría sobre las últimas 7 predicciones para reducir fluctuaciones, y (4) aplicación web con Streamlit capturando video del navegador y superponiendo etiquetas en tiempo real. Las pruebas confirmaron clasificación consistente de actividades de desplazamiento pero dificultades con transiciones sentarse/pararse, coherente con resultados offline.

Resultados

Después del procesamiento de los 126 videos (generados tanto por el equipo

como mediante intercambio colaborativo con otros grupos), se extrajeron para cada video las siguientes métricas agregadas:

- Velocidad promedio de la cadera
- Inclinación promedio de los hombros
- Ángulo promedio de rodilla
- Brillo del frame
- Movimiento inter-frame
- Duración y número de frames
- Etiqueta de acción asignada por nombre de archivo

El dataset final contenía

Clase	Muestras	Porcentaje
caminar_atras	19	15.08%
caminar_hacia_adelante	19	15.08%
girar_derecha	28	22.22%
pararse	26	20.63%
sentarse	34	26.98%

Se observa un leve desbalance, aunque no extremo, con predominancia de sentarse y girar_derecha. Por este motivo se utilizaron estrategias de compensación mediante class_weight='balanced'.

Validación Cruzada

Modelo	F1 promedio CV
Random Forest	**0.8282**
SVM	0.8251
XGBoost	0.8073

El **Random Forest** obtuvo el mejor desempeño global durante la validación.

Evaluación en el conjunto de prueba

Se usó un split estratificado 80/20. A continuación se resumen las métricas:

Random Forest

- Accuracy: **0.77**
- F1-score macro: **0.75**

Tabla de desempeño por clase:

Clase	Precision	Recall	F1
caminar_atras	1.00	0.50	0.67
caminar_hacia_adelante	0.60	0.75	0.67
girar_derecha	0.75	1.00	0.86
pararse	1.00	0.60	0.75
sentarse	0.75	0.86	0.80

SVM

- Accuracy: **0.81**
- F1-score macro: **0.79**

Tabla de desempeño por clase

Clase	Precision	Recall	F1
caminar_atras	0.75	0.75	0.75
caminar_hacia_adelante	0.75	0.75	0.75
girar_derecha	1.00	1.00	1.00

pararse	0.75	0.60	0.67
sentarse	0.75	0.86	0.80

XGBoost

- Accuracy: **0.73**
- F1-score macro: **0.70**

Tabla de desempeño por clase

Clase	Precision	Recall	F1
caminar_atras	0.67	0.50	0.57
caminar_hacia_adelante	0.50	0.50	0.50
girar_derecha	0.75	1.00	0.86
pararse	1.00	0.60	0.75
sentarse	0.75	0.86	0.80

Análisis de Resultados

Los resultados obtenidos muestran una diferencia importante entre el desempeño de los modelos durante la evaluación offline y su funcionamiento en tiempo real. Aunque las métricas numéricas sugerían buen rendimiento para varias actividades, las pruebas prácticas evidenciaron limitaciones que afectan la capacidad del sistema para generalizar correctamente.

1. Generalización real del modelo

En la evaluación offline (con métricas agregadas por video), los modelos parecían distinguir adecuadamente varias clases. Sin embargo, al integrarlos en el pipeline de tiempo real, el sistema mostró un comportamiento diferente:

- Sólo “caminar hacia adelante” y “caminar atrás” fueron detectados de forma consistente.
- Las actividades “sentarse”, “pararse” y especialmente “girar” no fueron reconocidas con fiabilidad.

Esto evidencia una brecha entre el desempeño medido con promedios por video y la dinámica real del movimiento cuando se analiza frame a frame.

2. Actividades correctamente clasificadas

Caminar hacia adelante y caminar atrás

Estas fueron las únicas actividades detectadas de forma estable en la aplicación en tiempo real. Las razones principales son:

- Son movimientos largos y repetitivos.
- Generan patrones consistentes en velocidad de cadera y movimiento inter-frame.
- La ventana deslizante de 15 frames captura suficiente información para caracterizar la marcha.
- MediaPipe rastrea estos landmarks con alta estabilidad cuando la persona está completamente visible.

Aunque hubo confusión ocasional entre ambas (especialmente cuando la persona caminaba muy lento o cerca de cámara), fueron las clases con mejor desempeño práctico.

3. Actividades con desempeño deficiente en tiempo real

Sentarse y pararse

El sistema intentó clasificarlas, pero con poca precisión y alta inestabilidad:

- En algunos casos las predicciones fluctuaban entre “caminar” y etiquetas incorrectas.
- La fase de transición dura pocos frames y no alcanza a llenar la ventana temporal requerida.
- Las variaciones del ángulo de rodilla no siempre son capturadas adecuadamente cuando la persona se sienta parcialmente o está fuera de cuadro.

En consecuencia, estas actividades fueron clasificadas solo en circunstancias muy específicas (movimientos lentos, bien centrados y completamente visibles).

Girar a la derecha

Esta actividad fue la más problemática y prácticamente **no se detectó en ningún momento** durante las pruebas reales.

Las causas principales son:

1. **MediaPipe pierde landmarks al girar el torso**, especialmente en rotaciones laterales donde partes del cuerpo dejan de ser visibles.
2. El sistema utiliza **características promedio**, pero un giro es un movimiento corto y abrupto que no se mantiene constante durante varios frames.
3. La ventana de 15 frames diluye el cambio súbito y lo promedia con estados anteriores y posteriores.
4. La información direccional no está explícita en landmarks 2D; se necesitaría orientación 3D del esqueleto o cálculos de rotación global.

Por estos motivos, aunque el modelo offline mostraba buenos resultados para esta clase, en inferencia real no fue posible reconocerla correctamente.

Conclusiones y Trabajo Futuro

El proyecto permitió diseñar e implementar un sistema funcional de reconocimiento de actividades humanas basado en MediaPipe Pose y modelos supervisados tradicionales. Se logró desarrollar una arquitectura completa que incluye recolección de datos, extracción de características cinemáticas, entrenamiento de modelos y despliegue en tiempo real mediante Streamlit.

Los resultados evidencian que el enfoque es adecuado para reconocer actividades continuas y prolongadas, como caminar hacia adelante y caminar atrás, que presentan patrones estables y fácilmente capturables por las características agregadas. Sin embargo, se identificó una brecha importante entre el desempeño offline y el funcionamiento real del sistema. Actividades de transición como sentarse y pararse, y especialmente movimientos con rotación del torso como girar, no pudieron ser reconocidas de forma fiable durante la inferencia en vivo. Esto demuestra que la representación basada en promedios y la ventana temporal utilizada no son suficientes para capturar la dinámica de movimientos cortos o complejos.

A partir de las limitaciones observadas, se identifican varias líneas de mejora que podrían aumentar significativamente la precisión y robustez del sistema:

1. Utilizar información 3D o estimación de orientación
Modelos como MediaPipe Holistic o pose estimators 3D permitirían detectar giros y

rotaciones que no pueden representarse correctamente con landmarks 2D.

2. Aumentar y equilibrar el dataset
Recolectar más ejemplos de sentarse, pararse y especialmente de giros, con variaciones en velocidad, ángulo y contexto, para mejorar la generalización.

3. Ajustar la ventana temporal y mejorar el smoothing
Explorar ventanas más amplias o adaptativas, así como métodos de estabilización más avanzados para reducir el ruido de landmarks.

4. Implementar mecanismos de detección de calidad
Agregar filtros para descartar frames con baja visibilidad, o para detectar cuando el cuerpo está parcialmente fuera de cuadro.

Con estas mejoras, el sistema podría evolucionar hacia un detector más robusto, capaz de manejar movimientos complejos y condiciones reales más desafiantes, acercándose a aplicaciones prácticas como monitoreo clínico, ergonomía laboral o análisis deportivo.

Bibliographic References:

1. J. W. Kim, S. Lee, K. Kim, et al., “Human Pose Estimation Using MediaPipe Pose and Fast Optimization Methods,” *Appl. Sci.*, vol. 13, no. 4, p. 2700, 2023. (MDPI)
2. A. A. K. Kumar, A. K. Singh, S. K. Singh, R. Kala, “Human Activity Recognition in Real-Time Environments Using Skeleton Joints,” *Int. J. Interact. Multimed. Artif. Intell.*, vol. 3, no. 7, pp. 62–70, Sept. 2016. (Dialnet)
3. W. Z. Tee, J. Seliya, R. Dave, and M. Vanamala, “A Close Look into Human Activity Recognition Models Using Deep Learning,” 2022. [Online]. Available: arXiv:2204.13589. (arXiv)
4. N. Sedaghati, “Application of Human Activity/Action Recognition: A Review,” *Multimedia Tools Appl.*, 2025. (SpringerLink)
5. A. A. R. Bsoul, “Human Activity Recognition Using Graph Structures and Deep Neural Networks,” *Comput. Intell.*, vol. 14, no. 1, p. 9, 2024. (MDPI)

Link del repositorio:
https://github.com/JuanCami009/video-annotation-system.git

Link video:
https://youtu.be/iGJfDw4XV7g