# Clasificación de Actividades Humanas en Tiempo Real: Del Alto Rendimiento Offline a los Desafíos del Despliegue en Producción

Juan Esteban Ruiz
*Departamento de Ingeniería de Sistemas*
*Universidad ICESI*
Cali, Colombia
juan.ruiz@u.icesi.edu.co

Juan David Quintero Peña
*Departamento de Ingeniería de Sistemas*
*Universidad ICESI*
Cali, Colombia
juan.quintero@u.icesi.edu.co

Tomás Quintero
*Departamento de Ingeniería de Sistemas*
*Universidad ICESI*
Cali, Colombia
tomas.quintero@u.icesi.edu.co

***Resumen*—El reconocimiento de actividades humanas (HAR) mediante visión por computadora presenta un potencial transformador para la rehabilitación física remota y el monitoreo de adultos mayores. Este trabajo presenta el desarrollo de un sistema de clasificación para cinco actividades básicas (caminar hacia adelante/atrás, girar, sentarse, ponerse de pie) utilizando un enfoque basado en Random Forest sobre características geométricas extraídas con MediaPipe. Aunque el modelo alcanzó un rendimiento excepcional en condiciones controladas (accuracy del 98.76 % en el conjunto de prueba y estabilidad validada mediante Bootstrap), el despliegue en tiempo real reveló un "gap" significativo de rendimiento, cayendo la precisión a un estimado del 40 % en entornos de producción no controlados. Este informe detalla la metodología rigurosa de entrenamiento, incluyendo estrategias de aumento de datos sin fugas de información, y analiza críticamente las discrepancias observadas entre las métricas offline y la realidad operativa, identificando la ausencia de contexto temporal y la variabilidad de los landmarks como causas raíz principales.**

***Index Terms*—Human Activity Recognition, MediaPipe, Random Forest, Despliegue en Tiempo Real, Machine Learning.**

## I. Introduction

El análisis automatizado del movimiento humano se ha convertido en una herramienta crucial en diversos dominios, desde el análisis deportivo hasta la vigilancia inteligente. Sin embargo, su impacto potencial es más crítico en el sector salud, específicamente en la rehabilitación física remota y el monitoreo de seguridad para una población mundial que envejece rápidamente. Se estima que para 2050, el 22 % de la población mundial superará los 65 años, incrementando la demanda de sistemas autónomos de detección de caídas y evaluación de movilidad.

Los métodos tradicionales de evaluación clínica son manuales, costosos y limitados por la disponibilidad de especialistas. Este proyecto busca desarrollar un sistema automatizado, accesible y de bajo costo, capaz de clasificar actividades fundamentales utilizando hardware estándar (webcams) en lugar de costosos sensores wearables o sistemas de captura de movimiento óptico.

El problema abordado es la clasificación supervisada de cinco actividades mutuamente excluyentes: caminar hacia la cámara, caminar de regreso, girar, sentarse y ponerse de pie. El desafío técnico no solo reside en la variabilidad interpersonal (altura, velocidad, complexión), sino en la necesidad de inferencia en tiempo real ($<$ 100ms de latencia) para aplicaciones interactivas.

Este informe documenta el ciclo completo de desarrollo bajo la metodología CRISP-DM, desde la recolección de datos y el entrenamiento de un modelo con un accuracy offline superior al 98 %, hasta los desafíos críticos encontrados durante el despliegue en un entorno de producción real, destacando la discrepancia entre las métricas de laboratorio y el rendimiento ïn the wild" [1].

## II. Theory

El desarrollo de un sistema de reconocimiento de actividades (HAR) robusto requiere una fundamentación teórica sólida en tres pilares: la representación vectorial del cuerpo humano, la invarianza geométrica y el manejo de desequilibrios en el espacio de características.

### *II-A. Representación del Espacio de Pose*

Utilizamos el framework MediaPipe Pose [2], el cual modela el cuerpo humano como un grafo cinemático topológico. A diferencia de los métodos basados en mapas de calor (Heatmaps) como OpenPose [3], MediaPipe infiere coordenadas de regresión directa, lo que permite una ejecución eficiente en CPU.

Para cada frame $t$, el modelo infiere un conjunto $L_t$ de 33 landmarks corporales:

$$L_t = \{\mathbf{p}_i \mid i \in [0, 32]\} \tag{1}$$

Donde cada punto $\mathbf{p}_i$ se define en un espacio tetradimensional:

$$\mathbf{p}_i = (x_i, y_i, z_i, v_i) \tag{2}$$

Siendo $(x_i, y_i)$ las coordenadas normalizadas en el plano de la imagen $[0, 1]$, $z_i$ la profundidad relativa al centro de la cadera (escala aproximada en metros), y $v_i \in [0, 1]$ la probabilidad de visibilidad del landmark.

### II-B. Ingeniería de Características Invariantes

Los landmarks crudos presentan alta variabilidad debido a la posición del sujeto en el encuadre y su distancia a la cámara. Para construir un clasificador robusto, es necesario transformar $L_t$ en un vector de características geométricas $\mathbf{f}_t$ que sea invariante a la escala y traslación.

*II-B1. Normalización por Escala del Torso:* Definimos una métrica de referencia $S_{ref}$ basada en la longitud del torso, la cual es anatómicamente estable durante el movimiento de extremidades:

$$S_{ref} = \left\| \frac{\mathbf{p}_{11} + \mathbf{p}_{12}}{2} - \frac{\mathbf{p}_{23} + \mathbf{p}_{24}}{2} \right\|_2 \tag{3}$$

Donde $\mathbf{p}_{11}, \mathbf{p}_{12}$ son los hombros y $\mathbf{p}_{23}, \mathbf{p}_{24}$ son las caderas. Cualquier distancia euclidiana $d_{raw}$ entre dos articulaciones se normaliza mediante:

$$d_{norm}(a, b) = \frac{\|\mathbf{p}_a - \mathbf{p}_b\|_2}{S_{ref}} \tag{4}$$

*II-B2. Cinemática Angular:* Los ángulos articulares son críticos para distinguir estados biomecánicos (ej. flexión de rodilla en "Sentarse"vs extensión en "De pie"). Calculamos el ángulo $\theta$ en una articulación central $B$ conectada a $A$ y $C$:

$$\theta = \arccos\left( \frac{\vec{v}_{BA} \cdot \vec{v}_{BC}}{\|\vec{v}_{BA}\| \|\vec{v}_{BC}\|} \right) \tag{5}$$

Donde $\vec{v}_{BA} = \mathbf{p}_A - \mathbf{p}_B$. Este cálculo se aplica bilateralmente a codos, rodillas y caderas.

### II-C. Sobre-muestreo Sintético (SMOTE)

El desbalance de clases es endémico en datasets propios. Utilizamos SMOTE [4] para generar muestras sintéticas en el espacio de características. Para una muestra minoritaria $\mathbf{x}_i$, se interpola con un vecino cercano $\mathbf{x}_{nn}$:

$$\mathbf{x}_{new} = \mathbf{x}_i + \lambda \cdot (\mathbf{x}_{nn} - \mathbf{x}_i), \quad \lambda \sim U(0, 1) \tag{6}$$

Esto densifica las fronteras de decisión sin el riesgo de sobreajuste asociado a la simple duplicación de datos.

## III. Methodology

La metodología se estructuró siguiendo el estándar CRISP-DM, con un énfasis riguroso en la integridad de los datos para evitar fugas de información (*data leakage*) entre las fases de entrenamiento y evaluación.

### III-A. Recolección y Curación del Dataset

Se construyó un dataset propietario con 90 videos (aprox. 1.2 horas de metraje) provenientes de 18 sujetos. Los criterios de inclusión garantizaron variabilidad en antropometría (alturas 1.58m - 1.83m) y ejecución. El análisis inicial mostró un desbalance severo: la clase "Girarrepresentaba solo el 14.6 % de los frames, versus el 28.6 % de Çaminar Hacia".

### III-B. Ingeniería de Características

Transformamos los 33 landmarks crudos en un vector de 83 características diseñadas para maximizar la separabilidad de clases. La Tabla I detalla la composición de este vector.

Tabla I
Definición del Espacio de Características (83 D)

| Tipo | Descripción |
|---|---|
| **Landmarks (64)** | Coordenadas $(x, y)$ normalizadas por la resolución de la imagen y coordenadas de profundidad relativa $z$ para 16 articulaciones clave (hombros, codos, muñecas, caderas, rodillas, tobillos, talones, puntas de pie). Se excluyeron los landmarks faciales detallados por su baja relevancia biomecánica. |
| **Distancias (8)** | Distancias euclidianas normalizadas por $S_{ref}$ (Ec. 3): Anchura de hombros, anchura de caderas, longitud de torso (izq/der), longitud de muslo (izq/der), longitud de pantorrilla (izq/der). |
| **Ángulos (4)** | Ángulos articulares (Ec. 5) para: Codo izquierdo, Codo derecho, Rodilla izquierda, Rodilla derecha. Rango $[0, \pi]$. |
| **Ratios (3)** | Proporciones biomecánicas adimensionales: Ratio hombro/cadera, Ratio torso/pierna, Aproximación de altura corporal. |
| **Centro de Masa (4)** | Coordenadas del centroide del tren superior e inferior relativos al centro del frame. |

### III-C. Pipeline de Preprocesamiento Estricto

Para evitar el *data leakage*, implementamos el siguiente flujo secuencial:

*III-C1. Split Estratificado Temprano:* La división de datos se realizó a nivel de *video completo*, asegurando que frames correlacionados de una misma secuencia no aparezcan en conjuntos diferentes.

$$D_{total} = D_{train}(70\,\%) \cup D_{val}(15\,\%) \cup D_{test}(15\,\%) \tag{7}$$

*III-C2. Balanceo y Transformación:* Se aplicó SMOTE **únicamente** sobre $D_{train}$, elevando las clases minoritarias hasta un factor de balance $\beta = 0{,}8$. Posteriormente, se ajustó un *StandardScaler* ($Z = \frac{x-\mu}{\sigma}$) solo con los datos de entrenamiento y se aplicó la transformación a validación y prueba.

*III-C3. Reducción de Dimensionalidad (PCA):* Aplicamos Análisis de Componentes Principales sobre el vector de 83 dimensiones. Seleccionamos los primeros $k = 16$ componentes que explican el 95.1 % de la varianza total:

$$\sum_{i=1}^{16} \lambda_i \approx 0{,}951 \cdot \sum_{j=1}^{83} \lambda_j \tag{8}$$

### III-D. Algoritmo de Despliegue en Tiempo Real

Para la fase de producción, desarrollamos un pipeline de inferencia optimizado para baja latencia. El proceso se formaliza en el Algoritmo III-D.

Pipeline de Inferencia en Tiempo Real

**Require:** Stream de video $V$, Modelo entrenado $M$, Transformador PCA $T_{pca}$, Escalador $S$
**Ensure:** Predicción de clase $C_{final}$ por frame

$B \leftarrow$ EmptyDeque($maxlen = 10$) {Buffer de suavizado}

**while** $V$ has frames **do**
  $F_t \leftarrow$ GetNextFrame($V$)
  $L_t \leftarrow$ MediaPipePose($F_t$)
  **if** $L_t$ is detected **then**
    $feat_{raw} \leftarrow$ ExtractGeometricFeatures($L_t$)
    $feat_{norm} \leftarrow S$.transform($feat_{raw}$)
    $feat_{pca} \leftarrow T_{pca}$.transform($feat_{norm}$)
    $prob_t \leftarrow M$.predict_proba($feat_{pca}$)
    $C_t \leftarrow$ argmax($prob_t$)
    $B$.append($C_t$)
    $C_{final} \leftarrow$ Mode($B$) {Voto mayoritario}
  **else**
    $C_{final} \leftarrow$ NoDetection
  **end if**
  **return** $C_{final}$
**end while**

## IV. Results

La evaluación se dividió en dos fases: validación cuantitativa offline sobre el conjunto de prueba retenido y validación cualitativa online en entorno real.

### *IV-A. Métricas de Rendimiento Offline*

El modelo Random Forest seleccionado alcanzó un **Accuracy global del 98.76 %** en el conjunto de prueba (Test Set, $n = 967$ frames). La Tabla II compara el rendimiento entre validación y prueba.

Tabla II
Comparación de Rendimiento (Validation vs Test)

| Métrica | Validation Set | Test Set | Δ |
|---|---|---|---|
| Accuracy | 98.60 % | **98.76 %** | +0.16 % |
| Macro F1 | 98.50 % | 98.76 % | +0.26 % |
| Weighted F1 | 98.60 % | 98.76 % | +0.16 % |
| Errores Totales | 14 / 967 | **12 / 967** | -2 |

Es notable que el rendimiento en Test superó ligeramente a Validation (+0.16 %), descartando overfitting. Para validar la significancia estadística, ejecutamos un análisis *Bootstrap* con 1,000 iteraciones, obteniendo un intervalo de confianza del 95 % de $[98{,}0\,\%, 99{,}4\,\%]$ y un coeficiente de variación de apenas 0.36 %.

### *IV-B. Desempeño Detallado por Clase*

La Tabla III desglosa las métricas por actividad. Se observa un comportamiento casi ideal en actividades cíclicas y distintivas como Çaminar de Regresoz "Girar". Las confusiones residuales se concentraron en las clases "Sentarsez "Ponerse de Pie", lo cual es biomecánicamente coherente dado que ambas comparten la misma trayectoria espacial (flexión de cadera/rodilla) diferenciándose únicamente por la dirección del vector de velocidad temporal, característica que el modelo estático captura con menor precisión.

Tabla III
Reporte de Clasificación por Actividad (Test Set)

| Actividad | Precision | Recall | F1-Score | Muestras |
|---|---|---|---|---|
| Caminar Hacia | 0.98 | 1.00 | 0.99 | 277 |
| Caminar Regreso | 1.00 | 1.00 | 1.00 | 195 |
| Girar | 1.00 | 0.99 | 1.00 | 141 |
| Ponerse de Pie | 0.98 | 0.98 | 0.98 | 166 |
| Sentarse | 0.98 | 0.96 | 0.97 | 188 |

### *IV-C. Comparación de Modelos: RF vs MLP*

Se evaluó comparativamente un Perceptrón Multicapa (MLP) frente al Random Forest seleccionado. Aunque el MLP alcanzó un accuracy marginalmente superior (98.97 % vs 98.76 %), el Random Forest demostró una velocidad de inferencia tres veces superior ($\sim 0{,}5$ms/frame vs $\sim 1{,}5$ms/frame). Dado el requisito estricto de tiempo real ($< 100$ms de latencia total incluyendo preprocesamiento), se priorizó la eficiencia del Random Forest.

## V. Results Analysis

El análisis de los resultados revela una dicotomía crítica entre el rendimiento teórico en el laboratorio y la operatividad en el mundo real.

### *V-A. El Fenómeno del "Gap Offline-Online"*

—

—

— —

[htbp] An“80“341lisis de Degradaci“80“363n: Laboratorio vs Producci“80“363n
#I|$l$|$c$|$c$|$c$|height

| Actividad | Test (Lab) | Prod (Est.) | Gap |
|---|---|---|---|
| Caminar Hacia | 99.6“80“045 | ∼85“80“045 | -15“80“045 |
| Caminar Regreso | 100.0“80“045 | ∼25“80“045 | -75“80“045 |
| Girar | 99.3“80“045 | ∼35“80“045 | -64“80“045 |

#I#I#$I$

#I

Mientras que Çaminar Hacia"mantuvo un rendimiento aceptable, actividades como Çaminar de Regresoz "Girar"sufrieron caídas catastróficas. Este fenómeno subraya que un alto *Test Accuracy* es una condición necesaria, pero no suficiente, para el despliegue exitoso.

### *V-B. Causas Raíz de la Falla en Producción*

Identificamos tres factores causales principales que explican esta discrepancia:

*V-B1. Ausencia de Contexto Temporal:* El modelo actual clasifica $P(Y|X_t)$, donde $X_t$ es el frame actual. Sin embargo, actividades como "Girar"se definen por la derivada temporal de la pose $\frac{dX}{dt}$. En producción, un frame estático aislado de un giro es geométricamente ambiguo y a menudo se confunde con Çaminar". La falta de memoria (estado oculto) impide al modelo entender la continuidad de la acción.

*V-B2. Sesgo de Perspectiva (Camera Angle Bias):* El dataset de entrenamiento consistía predominantemente en vistas frontales (0°).

$$P_{train}(\theta_{cam}) \approx \delta(\theta - 0^\circ) \tag{9}$$

En producción, cuando un usuario ejecuta Çaminar de Regreso", presenta una vista dorsal (180°). Aunque MediaPipe infiere los landmarks, la proyección geométrica resultante ($\mathbf{f}_{back}$) difiere significativamente de la frontal ($\mathbf{f}_{front}$) aprendida por el modelo, resultando en una clasificación errónea por *domain shift*.

*V-B3. Sensibilidad al Ruido de Detección:* En condiciones de laboratorio (iluminación controlada), la varianza del error de estimación de landmarks $\sigma^2_{noise}$ es baja. En producción (webcam doméstica, baja luz), $\sigma^2_{noise}$ aumenta, introduciendo ruido en el vector de características que empuja las muestras fuera de las fronteras de decisión aprendidas.

## VI. Conclusions and Future Work

Este trabajo presentó el ciclo de desarrollo completo de un sistema HAR, desde la recolección de datos hasta el despliegue. Logramos entrenar un modelo Random Forest con un rendimiento offline excepcional (98.76 %), validado mediante técnicas rigurosas de prevención de fugas de datos y análisis estadístico Bootstrap.

Sin embargo, la contribución más significativa de este estudio es la documentación empírica de las limitaciones de los modelos estáticos en entornos no controlados. Concluimos que la generalización ïn-the-wildrequiere más que un dataset balanceado; requiere arquitecturas que modelen explícitamente el tiempo y la variabilidad de perspectiva.

**Trabajo Futuro:** Para cerrar la brecha de rendimiento detectada, proponemos:

1. **Modelado Secuencial:** Implementar arquitecturas recurrentes como LSTM [5] que tomen como entrada una secuencia de frames $(X_{t-n}, ..., X_t)$ para capturar la dinámica temporal.
2. **Aumento de Datos Multi-vista:** Expandir el protocolo de recolección para incluir explícitamente ángulos de perfil (90°) y espalda (180°).
3. **Suavizado en Inferencia:** La implementación del buffer de votación (Algoritmo 1) mitigó parcialmente el ruido, pero se requieren filtros más sofisticados como Filtros de Kalman sobre los landmarks crudos antes de la extracción de características.

## Referencias


[1] X. Kong *et al.*, "Bridging the gap between training and inference for video super-resolution," *CVPR*, 2022.

[2] C. Lugaresi *et al.*, "Mediapipe: A framework for building perception pipelines," *arXiv preprint arXiv:1906.08172*, 2019.

[3] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," in *CVPR*, 2017.

[4] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: synthetic minority over-sampling technique," *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.

[5] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.