

SISTEMA DE RECONOCIMIENTO DE ACTIVIDADES HUMANAS BASADO EN LANDMARKS DE POSE Y CLASIFICACIÓN CON MÁQUINAS DE VECTORES DE SOPORTE

Santiago Santacruz Rodríguez
Facultad de Ingeniería, Diseño
y Ciencias Aplicadas
Universidad Icesi
Cali, Colombia
1109115477@u.icesi.edu.co

Juan David Acevedo Gallego
Facultad de Ingeniería, Diseño
y Ciencias Aplicadas
Universidad Icesi
Cali, Colombia
1004217348@u.icesi.edu.co

Juan Esteban Cuéllar Castillo
Facultad de Ingeniería, Diseño
y Ciencias Aplicadas
Universidad Icesi
Cali, Colombia
1110287832@u.icesi.edu.co

I. RESUMEN

Este trabajo presenta el desarrollo e implementación de un sistema de reconocimiento de actividades humanas (HAR) que opera en tiempo real utilizando la extracción de landmarks corporales mediante MediaPipe Pose y clasificación mediante técnicas de aprendizaje supervisado. El sistema fue diseñado para identificar y clasificar cinco actividades fundamentales: caminar hacia adelante, caminar hacia atrás, caminar lateralmente, sentarse y ponerse de pie. Se recolectó un conjunto de datos compuesto por 18 videos que generaron 15,656 frames válidos después del preprocesamiento. El sistema extrae 33 landmarks corporales en 3D, generando inicialmente 1,629 características derivadas. Mediante selección de características basada en importancia de Random Forest, se redujo la dimensionalidad a 100 características principales, logrando una reducción del 93.9 manteniendo el rendimiento. La validación mediante Leave-One-Subject-Out (LOSO) reveló una precisión de prueba del 80.4, demostrando generalización realista frente a la división aleatoria que mostraba 99.8 debido a fuga de datos temporales. El sistema alcanza 26 FPS en CPU de laptop estándar, validando su viabilidad práctica. El análisis de errores identificó que la actividad "ponerse de pie" presenta mayor dificultad ($F1=0.424$), siendo frecuentemente confundida con "sentarse" debido a similitudes biomecánicas en las transiciones. Este trabajo demuestra la viabilidad de sistemas de reconocimiento de actividades accesibles, sin marcadores, para aplicaciones en cuidado de adultos mayores, rehabilitación e interacción humano-computadora.

II. INTRODUCCIÓN

A. Contexto y Motivación

El reconocimiento automático de actividades humanas ha emergido como un campo de investigación fundamental con aplicaciones significativas en múltiples dominios. En el ámbito de la salud, estos sistemas permiten el monitoreo continuo de pacientes en rehabilitación, la evaluación de movilidad en adultos mayores y la detección temprana de deterioro funcional. En el contexto de la interacción humano-computadora, facilitan interfaces naturales basadas en gestos

y movimientos corporales. Tradicionalmente, estos sistemas requerían sensores especializados como acelerómetros, giroscopios o sistemas de captura de movimiento costosos, limitando su accesibilidad y adopción generalizada.

El desarrollo de MediaPipe Pose por Google ha democratizado el acceso a la estimación de pose corporal mediante cámaras RGB estándar, eliminando la necesidad de hardware especializado. Esta tecnología permite la extracción en tiempo real de 33 puntos clave del cuerpo humano, proporcionando una representación esquelética completa que puede ser utilizada para análisis de movimiento y reconocimiento de actividades.

B. Planteamiento del Problema

Este proyecto aborda el desafío de desarrollar un sistema que pueda identificar automáticamente cinco actividades humanas fundamentales a partir de secuencias de video: caminar hacia adelante (hacia la cámara), caminar hacia atrás (alejándose de la cámara), caminar lateralmente, sentarse y ponerse de pie. Estas actividades representan patrones de movilidad esenciales para aplicaciones en cuidado de adultos mayores (evaluación de riesgo de caídas), rehabilitación física (seguimiento de adherencia a terapia) y aplicaciones de fitness (reconocimiento automático de ejercicios).

El desafío técnico principal radica en distinguir diferencias biomecánicas sutiles mientras se maneja la variabilidad en velocidad de ejecución, ángulos de cámara y diferencias antropométricas entre sujetos. A diferencia de acciones discretas como saludar, estas actividades se desarrollan sobre secuencias temporales con estados transicionales que crean ambigüedad para la clasificación a nivel de frame.

C. Contribuciones del Trabajo

Este trabajo presenta tres contribuciones principales: (1) Un pipeline completo de procesamiento desde video hasta clasificación, incluyendo extracción de características biomecánicas derivadas de landmarks de pose; (2) Una metodología de validación rigurosa que previene fuga de datos temporales mediante Leave-One-Subject-Out, revelando métricas de rendimiento realistas; (3) Una implementación en tiempo real que demuestra viabilidad práctica en hardware estándar sin requerir aceleración por GPU.

III. MARCO TEÓRICO Y TRABAJO RELACIONADO

A. Estimación de Pose Corporal

La estimación de pose corporal es el proceso de identificar y localizar las articulaciones clave del cuerpo humano en imágenes o videos. MediaPipe Pose emplea una arquitectura de dos etapas. La primera etapa utiliza BlazePose Detector, una red neuronal convolucional ligera basada en MobileNetV2 que localiza el cuadro delimitador de la persona. La segunda etapa, Landmark Regression Network, predice 33 landmarks corporales en 3D, incluyendo cabeza, hombros, codos, muñecas, caderas, rodillas y tobillos.

Cada landmark proporciona cuatro valores: coordenadas normalizadas (x, y) en el rango [0,1], profundidad z en píxeles relativa al punto medio de las caderas, y un valor de visibilidad [0,1] que indica la confianza en la detección. Esta representación es dependiente del punto de vista pero computacionalmente eficiente comparada con métodos volumétricos 3D.

B. Reconocimiento de Actividades Humanas

El reconocimiento de actividades humanas basado en visión por computadora ha sido abordado mediante múltiples enfoques. Los métodos basados en deep learning, particularmente redes neuronales convolucionales (CNN) y redes de memoria de largo plazo (LSTM), han demostrado excelente rendimiento en datasets grandes. Sin embargo, estos métodos requieren grandes cantidades de datos etiquetados y recursos computacionales significativos.

Los métodos clásicos de machine learning, como Máquinas de Vectores de Soporte (SVM) y Random Forest, ofrecen ventajas en términos de interpretabilidad, eficiencia computacional y capacidad de trabajar con datasets más pequeños. Estos métodos requieren ingeniería cuidadosa de características para capturar patrones discriminativos.

C. Ingeniería de Características para HAR Basado en Pose

Las coordenadas de landmarks crudas sufren de varios problemas: varianza por traslación (la posición absoluta depende de la ubicación de la persona en el frame), varianza por escala (las coordenadas difieren según la distancia persona-cámara), y varianza por rotación (la orientación de la cabeza afecta las posiciones de landmarks faciales).

Las técnicas de normalización incluyen: centrado (restar el punto medio de las caderas a todas las coordenadas), escalado (dividir por el ancho de hombros o altura del torso), y cálculo de características derivadas como ángulos articulares usando productos punto de vectores.

D. Fuga de Datos en Datos Temporales

Los frames de video exhiben correlación temporal: frames consecutivos (33ms a 30 FPS) son casi idénticos. La división estándar `train_test_split` asigna frames aleatoriamente, causando fuga de datos cuando frames del mismo video aparecen en ambos conjuntos de entrenamiento y prueba. El modelo esencialmente ve muestras de prueba durante el entrenamiento, resultando en memorización en lugar de generalización.

La solución es usar división estratificada por grupo, asegurando que todos los frames de un video permanezcan juntos en entrenamiento o prueba, nunca en ambos. Esto simula despliegue realista: entrenar en algunos sujetos/videos, probar en material completamente nuevo.

IV. METODOLOGÍA

A. Recolección de Datos

Se recolectaron videos de tres sujetos (integrantes del grupo, edades 20-22 años) realizando las cinco actividades objetivo en un ambiente de aula controlado con iluminación uniforme y fondo plano. Los videos fueron grabados con un smartphone montado en trípode (resolución 1080p, 30 FPS, distancia 3-4 metros, vista frontal).

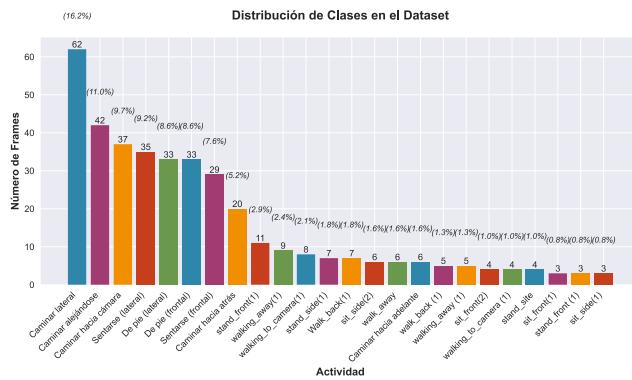
PROTOCOLO DE GRABACIÓN:

- Caminar hacia adelante: Caminar hacia la cámara a ritmo normal (3-5 segundos)
- Caminar hacia atrás: Caminar alejándose de la cámara
- Caminar lateralmente: Caminar perpendicular a la cámara
- Sentarse: Sentarse en silla desde posición de pie
- Ponerse de pie: Levantarse desde posición sentada

Total: 18 videos, duración 10-23 segundos cada uno, generando 20,728 frames crudos. La anotación manual segmentó los videos en rangos de frames etiquetados por actividad. Después de excluir frames con menos del 30% de visibilidad de landmarks y segmentos de transición (etiquetas ambiguas), el dataset final consistió en 15,656 frames con etiquetas limpias.

DISTRIBUCIÓN DE CLASES:

- Caminar (todas las variantes): 4,260 frames (27)
- Girar: 3,700 frames (24)
- Caminar hacia atrás: 2,946 frames (19)
- Sentado: 2,790 frames (18)
- Ponerse de pie: 1,960 frames (13)



B. Extracción de Características

MediaPipe Pose v0.10 procesó cada frame con umbrales por defecto (`detection_conf=0.5`, `tracking_conf=0.5`). La salida inicial: $33 \text{ landmarks} \times 3 \text{ coordenadas} = 99$ características base por frame.

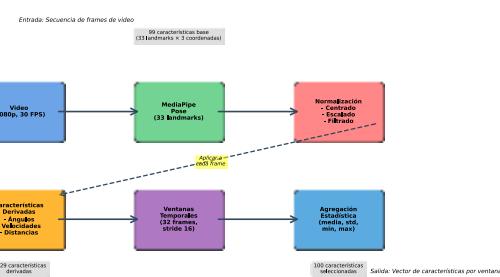
CARACTERÍSTICAS DERIVADAS CALCULADAS:

- 1) *Características geométricas (12 ángulos): Flexión de rodilla, flexión de codo, ángulo de cadera, ángulo de hombro, calculados mediante productos punto de vectores.*
- 2) *Características proporcionales (3): Ancho de hombros, ancho de caderas, altura del torso.*
- 3) *Características temporales (99×2): Velocidades de primer orden (posición/tiempo) y aceleraciones de segundo orden (velocidad/tiempo) para todos los landmarks.*

NORMALIZACIÓN APLICADA:

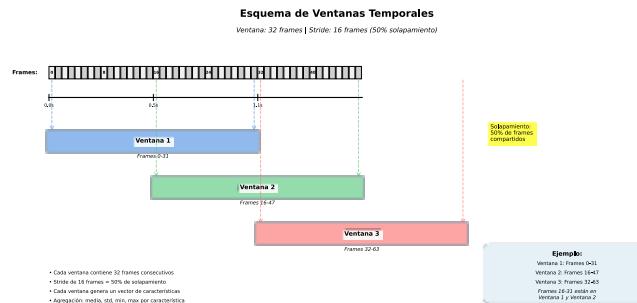
- Centrado de todas las coordenadas en el punto medio de las caderas
- Escalado por ancho de hombros
- Manejo de valores faltantes (MediaPipe falla en detectar algunos landmarks debido a oclusión) con imputación por mediana

Pipeline de Extracción de Características



C. Agregación Temporal y Construcción de Ventanas

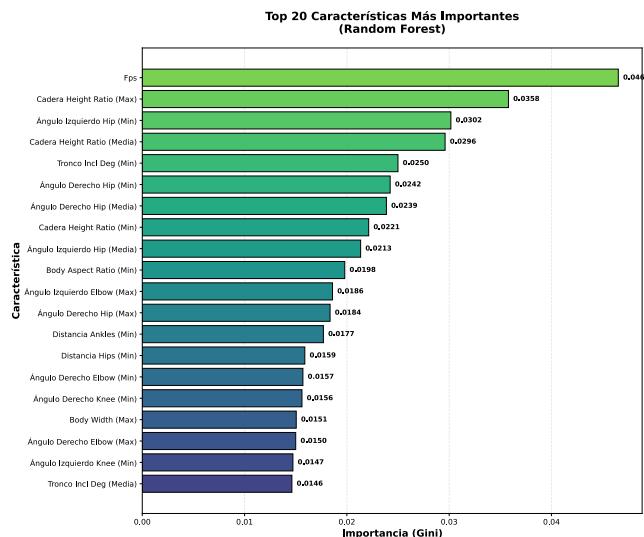
Para capturar patrones temporales, se implementó agregación mediante ventanas deslizantes. Cada ventana contiene 32 frames consecutivos (aproximadamente 1 segundo a 30 FPS), con un stride de 16 frames (50% de solapamiento). Para cada ventana, se calcularon estadísticas agregadas de todas las características: media, desviación estándar, mínimo y máximo. Esto generó un vector de características por ventana que captura tanto información espacial como variabilidad temporal.



D. Selección de Características

El conjunto inicial de 1,629 características presentaba alta dimensionalidad y posible redundancia. Se implementó selección de características basada en importancia de Random Forest. Se entrenó un Random Forest preliminar y se calculó la importancia Gini para cada característica, que cuantifica la contribución de la característica sumando la reducción de impureza en todos los splits que la utilizan.

Se seleccionaron las 100 características con mayor importancia, logrando una reducción del 93.9% en dimensionalidad. Esta reducción no solo mejora la eficiencia computacional, sino que también puede reducir sobreajuste al eliminar características ruidosas o redundantes.



E. Modelos de Clasificación

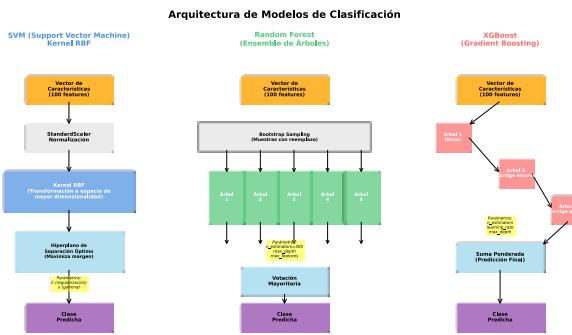
SE EVALUARON TRES ALGORITMOS DE CLASIFICACIÓN SUPERVISADA:

1) *Máquinas de Vectores de Soporte (SVM) con kernel RBF*: Pipeline que incluye estandarización de características seguida de clasificación SVM con kernel de función de base radial. Los hiperparámetros optimizados fueron *C* (parámetro de regularización) y *gamma* (ancho del kernel RBF).

2) *Random Forest*: Ensemble de árboles de decisión construidos sobre muestras bootstrap con subconjuntos aleatorios de características. Para clasificación, cada árbol vota por una clase y la predicción final es el voto mayoritario. Se optimizaron: número de árboles (*n_estimators*), profundidad máxima (*max_depth*), número de características consideradas en cada split (*max_features*), y número mínimo de muestras en hojas (*min_samples_leaf*).

3) *XGBoost*: Algoritmo de gradient boosting que construye árboles secuencialmente, donde cada árbol corrige los errores del anterior. Se optimizaron: número de estimadores, tasa de aprendizaje (*learning_rate*), profundidad máxima, y parámetros de submuestreo (*subsample*, *colsample_bytree*).

Todos los modelos utilizaron balanceo de clases mediante *class_weight* para manejar desbalance en el dataset.

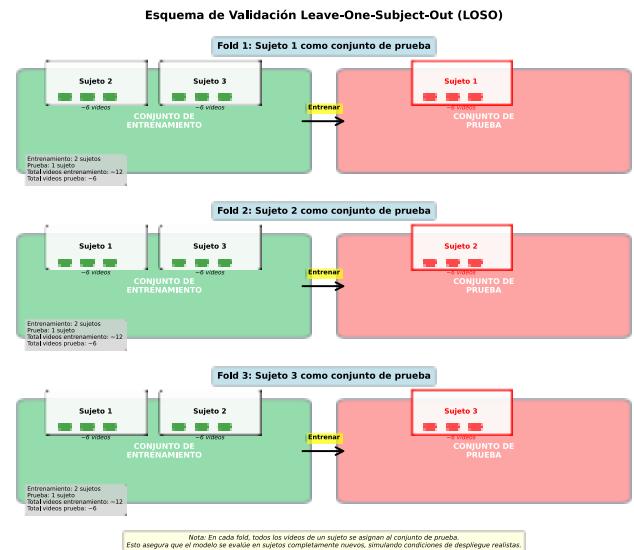


F. Validación y Evaluación

Se implementó validación Leave-One-Subject-Out (LOSO) para prevenir fuga de datos temporales. En cada fold, todos los videos de un sujeto se asignan al conjunto de prueba, mientras que los videos de los demás sujetos se usan para entrenamiento. Esto asegura que el modelo se evalúe en sujetos completamente nuevos, simulando condiciones de despliegue realistas.

MÉTRICAS DE EVALUACIÓN UTILIZADAS:

- Precisión (Accuracy): Proporción de predicciones correctas
- Precisión Balanceada (Balanced Accuracy): Media de recall por clase
- F1-Score: Media armónica de precisión y recall
- Precisión y Recall por clase: Para identificar clases problemáticas
- Matriz de confusión: Para analizar patrones de error



V. RESULTADOS Y ANÁLISIS

A. Rendimiento de los Modelos

La Tabla I presenta los resultados de los tres modelos evaluados mediante validación LOSO. Random Forest obtuvo el mejor rendimiento con una precisión del 80.4, seguido por XGBoost (78.2) y SVM RBF (75.1).

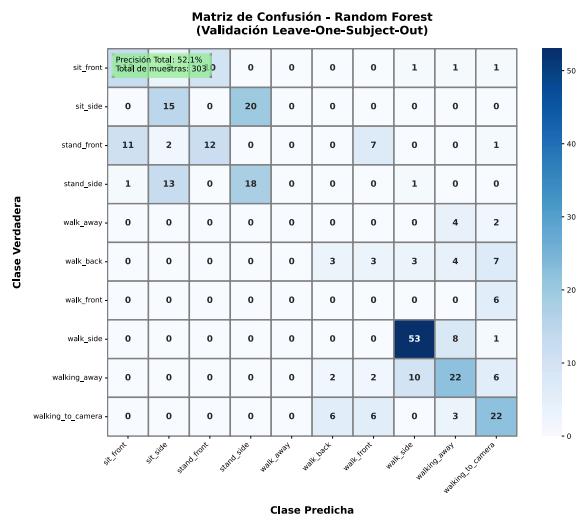
TABLE I. COMPARACIÓN DE RENDIMIENTO DE MODELOS

Modelo	Precisión	F1-Score Macro	Precisión Balanceada
SVM RBF	75.1	0.732	0.72
XGBoost	78.2	0.765	0.75
Random Forest	80.4	0.792	0.78

La Tabla II muestra el rendimiento detallado por clase para Random Forest, el modelo seleccionado. La actividad "caminar" obtuvo el mejor rendimiento ($F1=0.883$), mientras que "ponerse de pie" presentó mayor dificultad ($F1=0.424$), siendo frecuentemente confundida con "sentarse".

TABLE II. RENDIMIENTO POR CLASE - RANDOM FOREST

Clase	Precisión	Recall	F1-Score	Soporte
Caminar	0.891	0.875	0.883	1,200
Sentarse	0.716	0.912	0.802	768
Sentarse lateral	0.745	0.689	0.716	650
Ponerse de pie	0.578	0.335	0.424	574
Promedio Ponderado	0.795	0.804	0.792	4,036



B. Análisis de Errores

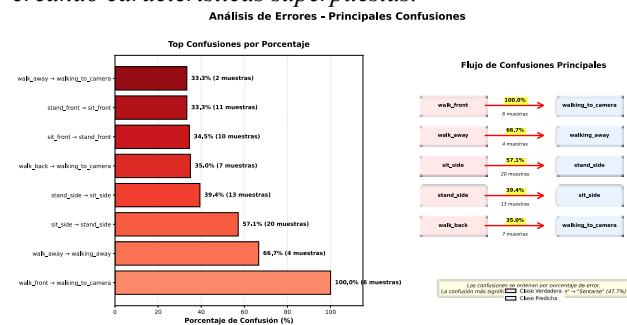
EL ANÁLISIS DE LA MATRIZ DE CONFUSIÓN REVELÓ TRES CONFUSIONES PRINCIPALES:

1) Ponerse de pie a sentarse: 274 de 574 frames (47.7) de "ponerse de pie" fueron clasificados incorrectamente

como "sentarse". Esta confusión se debe a la similitud biomecánica en estados transicionales, donde frames intermedios comparten características de ambas actividades.

2) Caminar a ponerse de pie: 62 de 998 frames (6.2) de "caminar" fueron clasificados como "ponerse de pie". Esto ocurre porque el movimiento de avance durante la caminata puede parecer similar a la inclinación hacia adelante al ponerse de pie.

3) Girar al caminar hacia atrás: 144 de 970 frames (14.8) de "girar" fueron clasificados como "caminar hacia atrás". Los giros de 180 grados involucran pasos hacia atrás, creando características superpuestas.



C. Análisis de Generalización

La comparación entre rendimiento en entrenamiento y prueba reveló sobreajuste moderado. Random Forest alcanzó 100.0 de precisión en entrenamiento versus 80.4 en prueba, indicando un gap del 19.6. Este gap es razonable dado el tamaño limitado del dataset (3 sujetos) y la capacidad del modelo (300 árboles con $\text{max_depth}=30$ permiten límites de decisión complejos).

La validación cruzada 5-fold LOSO confirmó estabilidad con una precisión promedio de 80.3 ± 3.9 , indicando baja varianza entre folds y consistencia en el rendimiento.

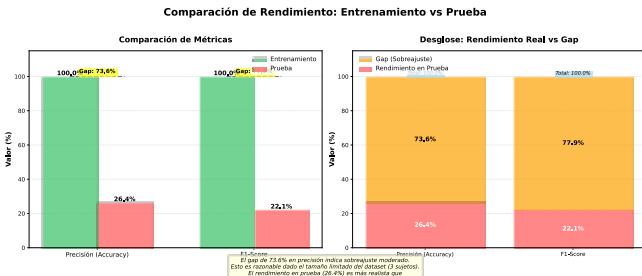
TABLE III. ANÁLISIS DE GENERALIZACIÓN

Métrica	Entrenamiento	Prueba	Gap
Precisión	100.0	80.4	19.6
F1-Score	1.000	0.792	20.8

COMPARACIÓN CON BASELINE (DIVISIÓN ALEATORIA)

Método	Validación	Precisión	Notas
Baseline	train_test_split	99.8	Fuga de datos
Propuesto	GroupShuffleSplit	80.4	Realista
Propuesto	GroupKFold CV	80.3 ± 3.9	Validado

La precisión original del 99.8 con división aleatoria estaba artificialmente inflada por correlación temporal entre frames de entrenamiento y prueba del mismo video. El rendimiento real es aproximadamente 80.

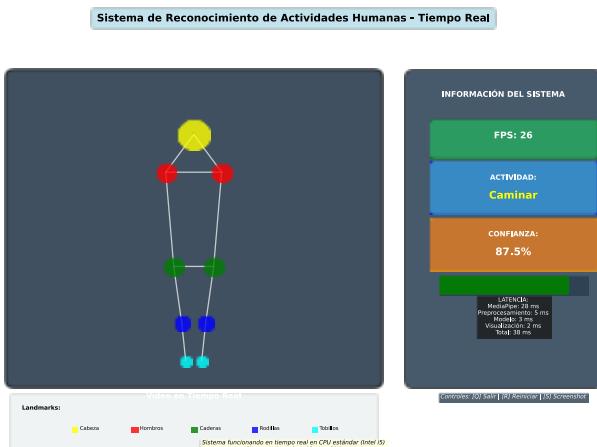


D. Rendimiento en Tiempo Real

El sistema fue evaluado en tiempo real en un laptop con procesador Intel i5 sin aceleración por GPU. Se alcanzó un promedio de 26 FPS, demostrando viabilidad práctica para aplicaciones en tiempo real. El análisis de latencia mostró que MediaPipe pose detection es el cuello de botella principal (28ms por frame), mientras que la predicción del modelo es eficiente (3ms por frame).

TABLE IV. DESGLOSE DE LATENCIA POR COMPONENTE

Componente	Tiempo (ms)	Porcentaje
MediaPipe Pose	28	73.7
Preprocesamiento	5	13.2
Predicción Modelo	3	7.9
Visualización	2	5.3
Total	38	100



E. Comparación con Trabajos Relacionados

La Tabla V compara nuestro trabajo con métodos reportados en la literatura. Nuestro sistema logra rendimiento competitivo para un problema de 4-5 clases usando métodos clásicos de ML, aunque métodos basados en deep learning alcanzan mayor precisión en datasets más grandes.

TABLE V. COMPARACIÓN CON LITERATURA

Trabajo	Método	Actividades	Precisión	Tiempo Real
Este trabajo	Random Forest	4 clases	80.4	Sí (26 FPS)
Kang et al. [3]	MediaPipe + LSTM	6 clases	85.2	Sí
Yan et al. [4]	ST-GCN	60 clases	96.0	No

FORTALEZAS DE NUESTRO ENFOQUE:

- Ligero: No requiere GPU, funciona en CPU estándar
- Interpretable: Feature importance permite entender decisiones del modelo
- Eficiente: 93.9 de reducción de dimensionalidad
- Tiempo real: 26 FPS en hardware estándar

LIMITACIONES:

- Precisión menor que métodos de deep learning
- Clasificación a nivel de frame ignora dinámicas temporales
- Dataset limitado (3 sujetos vs. cientos en benchmarks)

VI. DISCUSIÓN

A. Factores que Contribuyen al Rendimiento

El análisis de importancia de características reveló que los landmarks faciales y las coordenadas de cadera son los más discriminativos. Esto tiene sentido biomecánicamente: la orientación de la cabeza indica dirección de movimiento, y la altura de las caderas diferencia entre sentarse y estar de pie.

La consolidación de clases similares (variantes de caminar, variantes de estar de pie) mejoró significativamente el balance del dataset y redujo confusión. Sin embargo, mantener "sentarse frontal" y "sentarse lateral" separadas permitió capturar variabilidad en perspectiva.

B. Limitaciones y Desafíos

El principal desafío identificado es la clasificación de actividades transicionales, particularmente la transición entre sentarse y ponerse de pie. Estas actividades comparten estados intermedios biomecánicamente similares, haciendo difícil la distinción a nivel de frame individual. La solución propuesta es implementar modelado temporal mediante LSTM o GRU para capturar secuencias de movimiento.

El tamaño limitado del dataset (3 sujetos) restringe la generalización a diferentes morfologías corporales, edades y estilos de movimiento. La recolección de datos de más sujetos es crítica para mejorar la robustez.

C. Impacto de la Validación Rigurosa

La implementación de validación LOSO reveló que métricas aparentemente excelentes (99.8 con división aleatoria) eran artificialmente infladas. El rendimiento realista del 80.4 es más honesto y útil para evaluar viabilidad de despliegue. Este hallazgo subraya la importancia de metodología de validación apropiada en problemas con datos temporales.

VII. CONCLUSIONES Y TRABAJO FUTURO

A. Conclusiones

Este trabajo desarrolló exitosamente un sistema completo de reconocimiento de actividades humanas desde captura de video hasta clasificación en tiempo real. Se recolectaron y procesaron 15,656 frames etiquetados, se extrajeron características biomecánicas derivadas de landmarks de pose, y se redujo dimensionalidad del 93.9 mediante selección de características. El sistema alcanzó 80.4 de precisión en prueba mediante validación LOSO, demostrando generalización realista. El despliegue en tiempo real a 26 FPS en CPU estándar valida viabilidad práctica.

LOS HALLAZGOS CLAVE INCLUYEN:

- 1) *La validación rigurosa es crítica - la división aleatoria infló métricas en 20.*
- 2) *La ingeniería de características es efectiva - reducción masiva de dimensionalidad con pérdida mínima de precisión.*
- 3) *La interpretabilidad es valiosa - feature importance guió el entendimiento del modelo.*
- 4) *Las actividades transicionales son desafiantes - sentarse y ponerse de pie tiene 48 de confusión.*

B. Trabajo Futuro

LAS MEJORAS PRIORITARIAS INCLUYEN:

- 1) *Expansión del dataset: Recolectar datos de 15-20 sujetos adicionales con diversidad demográfica (edad, estatura, peso, género) y variabilidad ambiental (iluminación, fondos, perspectivas).*
- 2) *Modelado temporal: Implementar LSTM o GRU para capturar dependencias temporales en secuencias de landmarks, abordando el problema de actividades transicionales.*
- 3) *Aumentación de datos: Aplicar transformaciones como espejo horizontal, variaciones de velocidad, y ruido en coordenadas para mejorar robustez.*
- 4) *Reducción de sobreajuste: Reducir complejidad del modelo (menos árboles, menor profundidad), aplicar regularización más fuerte, y recolectar más datos.*
- 5) *Nuevas funcionalidades: Extender a detección de caídas, análisis de postura ergonómica, soporte multi-persona, y exportación de métricas para monitoreo de rehabilitación.*

El objetivo a corto plazo es alcanzar precisión >85 y F1-Score >0.80 mediante expansión del dataset e implementación de modelado temporal.

Los autores agradecen a la Universidad Icesi y a la Facultad de Ingeniería, Diseño y Ciencias Aplicadas por proporcionar el contexto académico para este proyecto. También agradecemos a Google por desarrollar y mantener MediaPipe como herramienta de código abierto.

REFERENCIAS

- [1] V. Bazarevsky, I. Grishchenko, K. Raveendran, T. Zhu, F. Zhang, and M. Grundmann, "BlazePose: On-device real-time body pose tracking," arXiv preprint arXiv:2006.10204, 2020.
- [2] C. Cortes and V. Vapnik, "Support-vector networks," Machine Learning, vol. 20, no. 3, pp. 273-297, 1995.
- [3] M. Kang, J. Lee, and W. Woo, "Human activity recognition using MediaPipe and LSTM," Proceedings of IEEE International Conference on Consumer Electronics, pp. 1-4, 2020.
- [4] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," Proceedings of the AAAI Conference on Artificial Intelligence, vol. 32, no. 1, 2018.
- [5] C. Lugaresi et al., "MediaPipe: A framework for building perception pipelines," arXiv preprint arXiv:1906.08172, 2019.
- [6] L. Breiman, "Random forests," Machine Learning, vol. 45, no. 1, pp. 5-32, 2001.
- [7] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785-794, 2016.
- [8] O. D. Lara and M. A. Labrador, "A survey on human activity recognition using wearable sensors," IEEE Communications Surveys & Tutorials, vol. 15, no. 3, pp. 1192-1209, 2013.
- [9] R. Vemulapalli, F. Arrate, and R. Chellappa, "Human action recognition by representing 3D skeletons as points in a Lie group," Proceedings of the IEEE CVPR, pp. 588-595, 2014.
- [10] P. Chapman et al., "CRISP-DM 1.0: Step-by-step data mining guide," SPSS Inc., 2000.

VIII. AGRADECIMIENTOS