Sistema de Reconocimiento de Actividades Humanas mediante Análisis de Pose con Deep Learning: Un Enfoque Basado en LSTM Bidireccional

Kevin Steven Nieto Curaca
Código: A00395466
Universidad ICESI
Cali, Valle del Cauca, Colombia
kevin.nieto@u.icesi.edu.co

Ricardo Urbina Ospina
Código: A00395489
Universidad ICESI
Cali, Valle del Cauca, Colombia
ricardo.urbina@u.icesi.edu.co

Julian Mendoza
Código: A00395237
Universidad ICESI
Cali, Valle del Cauca, Colombia
julian.mendoza@u.icesi.edu.co

***Abstract*—El reconocimiento de actividades humanas (HAR) representa un desafío fundamental en visión por computadora con aplicaciones críticas en salud, seguridad y análisis deportivo. Este estudio presenta un sistema integral para la clasificación de cinco actividades humanas fundamentales: caminar hacia adelante, caminar hacia atrás, girar, sentarse y ponerse de pie. Utilizando MediaPipe Pose para la extracción de landmarks corporales, se diseñó un pipeline completo que incluye feature engineering invariante a la fisiología del sujeto, normalización robusta y un modelo LSTM bidireccional. El dataset final comprende 1,468 secuencias temporales de 30 frames con características optimizadas. Los resultados preliminares en la clasificación binaria walk_forward vs walk_backward demuestran un desempeño excepcional con 100% de accuracy en validación y test, aunque se identifican áreas de mejora relacionadas con la diversidad del dataset. Este trabajo establece una base metodológica sólida para sistemas de análisis de movimiento humano con potencial aplicación en rehabilitación física y análisis biomecánico.**

***Index Terms*—Human Activity Recognition, Deep Learning, LSTM, MediaPipe, Pose Estimation, Feature Engineering, Computer Vision**

I. Introducción

El reconocimiento automático de actividades humanas (Human Activity Recognition - HAR) ha experimentado avances significativos en los últimos años gracias al desarrollo de algoritmos de deep learning y técnicas de estimación de pose en tiempo real. La capacidad de identificar y clasificar automáticamente movimientos humanos tiene aplicaciones transformadoras en múltiples dominios: sistemas de monitoreo de salud para adultos mayores, análisis biomecánico deportivo, interfaces humano-computadora naturales, y sistemas de seguridad inteligentes.

A. Planteamiento del Problema

La clasificación de actividades humanas a partir de video presenta desafíos técnicos fundamentales:

- **Variabilidad Fisiológica**: Las personas presentan diferencias significativas en altura, complexión, proporciones corporales y estilo de movimiento. Un sistema robusto debe ser invariante a estas características individuales.
- **Dependencias Temporales**: Las actividades humanas son inherentemente secuenciales. Un modelo efectivo debe capturar no solo la postura en un instante dado, sino también la evolución temporal del movimiento.
- **Invariancia a Condiciones de Captura**: El sistema debe funcionar independientemente de la distancia a la cámara, el ángulo de captura y las condiciones de iluminación.
- **Generalización entre Sujetos**: El modelo debe reconocer actividades realizadas por personas no vistas durante el entrenamiento.

B. Objetivos de Investigación

Este proyecto persigue los siguientes objetivos específicos:

1) Diseñar un pipeline completo de reconocimiento de actividades, desde la captura de video hasta la clasificación final.
2) Implementar una estrategia de feature engineering que genere descriptores invariantes a las características fisiológicas del sujeto.
3) Desarrollar un modelo de clasificación basado en LSTM optimizado para secuencias temporales de pose.
4) Evaluar el desempeño del sistema en cinco actividades fundamentales: caminar hacia adelante, caminar hacia atrás, girar, sentarse y ponerse de pie.
5) Identificar limitaciones del enfoque actual y proponer mejoras para trabajo futuro.

C. Contribuciones Principales

Las contribuciones clave de este trabajo incluyen:

- Un conjunto de características diseñadas específicamente para ser invariantes a la altura y proporciones corporales del sujeto.
- Una metodología rigurosa de feature selection que reduce la dimensionalidad de 50 a 39 features sin pérdida de información discriminativa.

- Un análisis exhaustivo del impacto del desbalance de clases y la homogeneidad del dataset en el rendimiento del modelo.
- Un framework reproducible implementado siguiendo las mejores prácticas de ingeniería de software.

II. Trabajo Relacionado

El reconocimiento de actividades humanas ha sido abordado desde múltiples perspectivas en la literatura. Los enfoques tradicionales basados en sensores inerciales (IMU) [1] ofrecen alta precisión pero requieren hardware especializado. Los métodos basados en visión por computadora se dividen principalmente en dos categorías: basados en apariencia (appearance-based) y basados en pose (pose-based).

A. Métodos Basados en Apariencia

Los enfoques de apariencia procesan directamente imágenes RGB o flujo óptico. Las redes convolucionales 3D (C3D) [2] y las Two-Stream Networks [3] han demostrado resultados estado del arte en datasets como UCF-101 y Kinetics. Sin embargo, estos métodos son computacionalmente costosos y sensibles a variaciones en el fondo y la iluminación.

B. Métodos Basados en Pose

Los métodos basados en pose estiman primero los keypoints del cuerpo humano y posteriormente clasifican la actividad. Esta separación de etapas ofrece ventajas significativas: invariancia al fondo, menor costo computacional y mayor interpretabilidad. OpenPose [4] y MediaPipe [5] son frameworks ampliamente utilizados para la estimación de pose en tiempo real.

Para la clasificación de secuencias de pose, las arquitecturas recurrentes como LSTM [6] y GRU han sido las opciones predilectas por su capacidad para modelar dependencias temporales de largo alcance. Trabajos recientes exploran también Transformers temporales [7], aunque a costa de mayor complejidad y requerimientos de datos.

C. Diferenciación del Trabajo Actual

Este estudio se distingue por:

1) Un énfasis explícito en el diseño de features invariantes a la fisiología del sujeto, utilizando ratios y ángulos normalizados en lugar de coordenadas absolutas.
2) Un análisis detallado del impacto de diferentes grupos de features (ángulos, distancias, ratios, direcciones, velocidades, temporales) en el desempeño del modelo.
3) Una metodología rigurosa de prevención de overfitting mediante feature selection basada en varianza e importancia discriminativa.

III. Metodología

El pipeline del sistema se estructura en seis fases principales, cada una diseñada para garantizar la robustez y reproducibilidad del proceso.

A. Adquisición y Organización de Datos

1) Dataset Inicial: Se recolectó un dataset de 53 videos de dos actividades: caminar hacia adelante (28 videos) y caminar hacia atrás (25 videos). Los videos fueron capturados con las siguientes especificaciones:

- Resolución: Mínimo 720p (1280x720)
- Frame rate: 30 fps
- Duración promedio: 4.2 segundos (126 frames)
- Sujetos: Múltiples personas con diferentes características físicas
- Condiciones: Iluminación controlada, fondo estático

2) Estructura de Directorios: El dataset se organizó siguiendo una estructura jerárquica que facilita el procesamiento automatizado:

```
data/
 source-1/
     raw/
         sit_down/
         stand_up/
         turning/
         walk_forward/
         walk_backward/
```

B. Extracción de Landmarks con MediaPipe

MediaPipe Pose [5] se utilizó para extraer 33 landmarks corporales en cada frame. Cada landmark se representa como un vector de 4 dimensiones: $(x, y, z, \text{visibility})$, donde:

- (x, y): Coordenadas normalizadas en el plano de la imagen $[0, 1]$
- z: Profundidad relativa
- visibility: Confianza de la detección $[0, 1]$

1) Landmarks de Interés: De los 33 landmarks, se seleccionaron 14 puntos clave relevantes para las actividades objetivo:

- **Superior**: Nariz (0), hombros (11, 12), codos (13, 14), muñecas (15, 16)
- **Inferior**: Caderas (23, 24), rodillas (25, 26), tobillos (27, 28), talones (29, 30), dedos del pie (31, 32)

2) Pipeline de Extracción: El proceso de extracción se implementó con las siguientes consideraciones técnicas:

1) **Resampling Temporal**: Todos los videos se normalizaron a 30 fps para consistencia.
2) **Manejo de Frames Inválidos**: Frames donde MediaPipe no detectó pose se rellenaron con arrays de ceros.
3) **Persistencia**: Los landmarks se guardaron en formato NumPy (.npy) para procesamiento eficiente posterior.

Resultados de Extracción:

- Videos procesados exitosamente: 53/53 (100%)
- Frames totales extraídos: 6,683
- Visibility promedio: 91.8% (walk_forward), 88.6% (walk_backward)

C. Feature Engineering

Esta fase representa el núcleo metodológico del proyecto. El objetivo fue transformar los landmarks crudos en descriptores de alto nivel invariantes a la fisiología del sujeto.

1) Normalización por Altura: Para cada frame, todos los landmarks se normalizaron dividiendo por la altura estimada de la persona:

$$h = \|\mathbf{p}_{\text{nose}} - \mathbf{p}_{\text{hip\_center}}\|_2 \tag{1}$$

$$\mathbf{p}_{\text{norm}} = \frac{\mathbf{p}}{h} \tag{2}$$

Donde $\mathbf{p}_{\text{hip\_center}} = \frac{\mathbf{p}_{\text{left\_hip}} + \mathbf{p}_{\text{right\_hip}}}{2}$

2) Grupos de Features: Se diseñaron 50 features agrupadas en 6 categorías:

1. Ángulos Articulares (12 features)

Ángulos en grados calculados usando la ley del coseno entre tres puntos:

$$\theta = \arccos\left(\frac{\mathbf{v}_1 \cdot \mathbf{v}_2}{\|\mathbf{v}_1\|\|\mathbf{v}_2\|}\right) \times \frac{180}{\pi} \tag{3}$$

Ejemplos: ángulo de rodilla (cadera-rodilla-tobillo), ángulo de cadera (hombro-cadera-rodilla).

2. Distancias Normalizadas (8 features)

Distancias euclidianas entre pares de landmarks, ya normalizadas por altura:

- Stride length: Distancia entre tobillos
- Arm swing: Distancia muñeca-cadera
- Torso length: Distancia centro_hombros-centro_caderas

3. Ratios Corporales (6 features)

Proporciones que son inherentemente invariantes a escala:

$$\text{Leg extension ratio} = \frac{d(\text{knee}, \text{ankle})}{d(\text{hip}, \text{knee})} \tag{4}$$

4. Vectores de Dirección (6 features)

Componentes normalizadas que indican orientación:

- Movement direction Z: Proyección en eje de profundidad
- Body orientation: Ángulo del vector hombros-caderas
- Lean forward/backward: Inclinación del tronco

5. Velocidades (10 features)

Derivadas temporales de posición (diferencia entre frames consecutivos):

$$v_t = p_t - p_{t-1} \tag{5}$$

Incluye velocidades lineales (tobillos, caderas) y angulares (rodillas).

6. Features Temporales (8 features)

Características calculadas sobre ventanas de frames:

- Cadence: Frecuencia de pasos detectada con `find_peaks`
- Gait phase: Fase del ciclo de marcha normalizada [0, 1]
- Contact flags: Indicadores binarios de contacto con el suelo
- Acceleration/Jerk: Derivadas de segundo y tercer orden

3) Resultados del Feature Engineering:

- Features generadas: 50 por frame
- Archivos procesados: 53 videos
- Total de frames con features: 6,683
- Tiempo de procesamiento: 1.8 segundos (promedio 28 videos/seg)

D. Feature Selection y Análisis de Correlación

1) Análisis Estadístico: Se realizó un análisis exhaustivo de las 50 features para identificar:

1) **Features con varianza nula** ($\sigma^2 < 10^{-6}$): 4 features eliminadas
   - angle_left_hip_frontal, angle_right_hip_frontal
   - step_length_ratio, symmetry_index
2) **Features altamente correlacionadas** ($|\rho| > 0.9$): Análisis de matriz de correlación reveló redundancias esperables entre features bilaterales (izquierda/derecha).
3) **Features discriminativas**: Análisis de diferencia de medias entre clases identificó las 20 features más relevantes.

2) Selección Final: Basándose en tres criterios (varianza, correlación, poder discriminativo), se eliminaron 11 features, resultando en un conjunto final de **39 features**:

- Ángulos: 10/12 conservadas
- Distancias: 8/8 conservadas
- Ratios: 4/6 conservadas
- Direcciones: 6/6 conservadas
- Velocidades: 7/10 conservadas
- Temporales: 4/8 conservadas

Features más discriminativas (top 5):

1) movement_direction_z (direccional)
2) body_orientation (direccional)
3) angle_right_knee (angular)
4) velocity_left_ankle_z (velocidad)
5) angle_left_hip (angular)

E. Construcción del Dataset Final

1) Generación de Ventanas Temporales: Se aplicó una técnica de sliding window para convertir las secuencias de features en muestras de tamaño fijo:

- **Window size**: 30 frames ($\sim$1 segundo a 30 fps)
- **Stride**: 15 frames (overlap del 50%)

Para un video de N frames, el número de ventanas generadas es:

$$n_{\text{windows}} = \left\lfloor \frac{N - W}{S} \right\rfloor + 1 \tag{6}$$

Donde $W = 30$ (window size) y $S = 15$ (stride).

2) Data Augmentation: Se aplicaron dos técnicas de augmentation:

1. Horizontal Flip (Espejo)

Intercambio sistemático de features left/right:

```
angle_left_knee <-> angle_right_knee
velocity_left_ankle <-> velocity_right_ankle
```

Factor de aumento: $2\times$

2. Speed Augmentation (Resampling Temporal)

Interpolación de la secuencia a diferentes velocidades:

- $0.8\times$ (20% más lento)
- $1.0\times$ (velocidad original)
- $1.2\times$ (20% más rápido)

Factor de aumento: $3\times$

Factor total de augmentation: $2 \times 3 = 6\times$ (aunque en práctica se aplicó selectivamente)

3) Normalización: Se aplicó StandardScaler (z-score normalization) sobre el conjunto de entrenamiento:

$$x_{\text{norm}} = \frac{x - \mu_{\text{train}}}{\sigma_{\text{train}}} \tag{7}$$

Crítico: Los parámetros μ y σ se calcularon solo sobre el conjunto de entrenamiento y se aplicaron a validación y test.

4) Split Estratificado: División realizada a nivel de video (no de ventanas) para evitar data leakage:

- **Train**: 70% (37 videos) $\rightarrow$ 1,020 ventanas
- **Validation**: 15% (8 videos) $\rightarrow$ 220 ventanas
- **Test**: 15% (8 videos) $\rightarrow$ 228 ventanas

Balance de clases mantenido en cada split ($\sim$50/50).

Shape final del dataset:

- X_{train}: $(1020, 30, 39)$
- X_{val}: $(220, 30, 39)$
- X_{test}: $(228, 30, 39)$

IV. Sit Down

A. Características Específicas de la Actividad

La actividad de sentarse representa una transición compleja que involucra la coordinación de múltiples articulaciones y el desplazamiento controlado del centro de masa corporal. Esta acción se caracteriza por una reducción progresiva de los ángulos articulares de las extremidades inferiores, acompañada de un descenso vertical del tronco.

Para la caracterización de esta actividad mediante MediaPipe Pose, se han identificado cuatro características fundamentales que capturan tanto el estado final como la dinámica de la transición.

1) Ángulo de Rodilla (θ_{knee}): El ángulo de rodilla se define como el ángulo formado por tres puntos anatómicos: cadera ($\mathbf{p}_{hip}$), rodilla ($\mathbf{p}_{knee}$) y tobillo ($\mathbf{p}_{ankle}$). Matemáticamente, se calcula mediante:

$$\mathbf{v}_1 = \mathbf{p}_{hip} - \mathbf{p}_{knee}, \quad \mathbf{v}_2 = \mathbf{p}_{ankle} - \mathbf{p}_{knee} \tag{8}$$

$$\theta_{knee} = \arccos\left(\frac{\mathbf{v}_1 \cdot \mathbf{v}_2}{\|\mathbf{v}_1\|\|\mathbf{v}_2\|}\right) \times \frac{180}{\pi} \tag{9}$$

Durante la acción de sentarse, este ángulo experimenta una disminución desde valores superiores a 160 (posición de pie) hasta valores típicamente inferiores a 130 (posición sentada). Esta característica resulta fundamental pues refleja directamente la flexión de la articulación de la rodilla, que constituye uno de los movimientos primarios en esta transición.

2) Ángulo de Cadera (θ_{hip}): El ángulo de cadera cuantifica la flexión del tronco respecto a las extremidades inferiores. Se define mediante tres puntos: el centro de los hombros ($\mathbf{p}_{shoulder}$), la cadera ($\mathbf{p}_{hip}$) y la rodilla ($\mathbf{p}_{knee}$):

$$\mathbf{v}_3 = \mathbf{p}_{shoulder} - \mathbf{p}_{hip}, \quad \mathbf{v}_4 = \mathbf{p}_{knee} - \mathbf{p}_{hip} \tag{10}$$

$$\theta_{hip} = \arccos\left(\frac{\mathbf{v}_3 \cdot \mathbf{v}_4}{\|\mathbf{v}_3\|\|\mathbf{v}_4\|}\right) \times \frac{180}{\pi} \tag{11}$$

Este ángulo complementa la información proporcionada por la rodilla, capturando la inclinación del tronco. La combinación de ambos ángulos permite discriminar con alta precisión entre las posturas de pie y sentado, eliminando ambigüedades que podrían surgir al considerar una sola articulación.

3) Velocidad Angular de Rodilla (ω_{knee}): La velocidad angular de rodilla captura la dinámica temporal de la transición. Se define como la diferencia del ángulo de rodilla entre frames consecutivos:

$$\omega_{knee}(t) = \theta_{knee}(t) - \theta_{knee}(t-1) \quad \text{[grados/frame]} \tag{12}$$

Durante la acción de sentarse, ω_{knee} presenta valores negativos consistentemente, indicando una reducción progresiva del ángulo articular. Esta característica es crucial para diferenciar el estado estático (sentado) de la acción dinámica (sentándose), ya que en estado estático $\omega_{knee} \approx 0$.

4) Velocidad Angular de Cadera (ω_{hip}): De manera análoga a la rodilla, la velocidad angular de cadera se calcula como:

$$\omega_{hip}(t) = \theta_{hip}(t) - \theta_{hip}(t-1) \quad \text{[grados/frame]} \tag{13}$$

Esta característica valida y refuerza la detección de la transición, proporcionando redundancia robusta ante posibles oclusiones o errores de detección en puntos individuales. La correlación entre ω_{knee} y ω_{hip} durante el movimiento permite identificar con alta confiabilidad la acción de sentarse.

5) Justificación de la Suficiencia del Conjunto de Características: El conjunto reducido de cuatro características (θ_{knee}, θ_{hip}, ω_{knee}, ω_{hip}) resulta suficiente para la clasificación de esta actividad debido a tres factores fundamentales.

Primero, las características seleccionadas capturan los dos aspectos esenciales del movimiento: el estado postural mediante los ángulos estáticos y la dinámica de transición mediante las velocidades angulares. Esta dualidad permite no solo identificar si una persona está sentada, sino también detectar el momento preciso en que ocurre la transición.

Segundo, la biomecánica del movimiento de sentarse está inherentemente determinada por estas articulaciones específicas. La física del movimiento dicta que la reducción coordinada de los ángulos de cadera y rodilla es necesaria y suficiente para lograr la postura sentada desde la posición de pie, por lo que no se requieren características adicionales de otras articulaciones.

Tercero, la alta separabilidad de las clases en el espacio de características reduce la necesidad de dimensionalidad adicional. Los valores de θ_{knee} y θ_{hip} para las posturas de pie y sentado presentan rangos claramente diferenciados (típicamente > 160 vs < 130), lo que facilita la clasificación con modelos simples y evita el riesgo de sobreajuste que podría surgir con un mayor número de características.

V. Stand Up

A. Características Específicas de la Actividad

La actividad de ponerse de pie constituye la transición inversa a sentarse, caracterizada por la extensión coordinada de las articulaciones de cadera y rodilla, junto con el desplazamiento ascendente del centro de masa. Desde una perspectiva biomecánica, esta acción requiere mayor esfuerzo muscular que sentarse, dado que implica trabajar contra la gravedad.

Las características empleadas para modelar esta actividad son idénticas en naturaleza a las utilizadas para sentarse, pero presentan patrones temporales y rangos de valores opuestos, lo que permite su diferenciación mediante el mismo conjunto de features.

1) Patrones Distintivos de Ángulos Articulares: Los ángulos de rodilla y cadera, definidos mediante las ecuaciones (2) y (4), exhiben durante la acción de ponerse de pie un comportamiento característico: transitan desde valores bajos (< 130) correspondientes a la posición sentada, hacia valores superiores a 160 propios de la postura erguida.

La trayectoria angular presenta típicamente una fase inicial de aceleración, donde los ángulos incrementan rápidamente, seguida de una fase de estabilización al alcanzar la postura de pie. Esta firma temporal permite distinguir la acción de ponerse de pie de otros movimientos que también involucren extensión articular.

2) Velocidades Angulares como Discriminantes Direccionales: Las velocidades angulares ω_{knee} y ω_{hip}, calculadas según las ecuaciones (5) y (6), constituyen el elemento discriminante fundamental entre sentarse y ponerse de pie. Durante la acción de stand up, ambas velocidades presentan valores consistentemente positivos:

$$\omega_{knee}(t) > 0, \quad \omega_{hip}(t) > 0 \quad \text{durante transición} \tag{14}$$

Esta característica direccional proporciona una señal inequívoca de la naturaleza de la transición. Mientras que en sentarse las velocidades son negativas (reducción angular), en ponerse de pie son positivas (incremento angular), permitiendo una clasificación binaria robusta basada únicamente en el signo de estas características.

3) Magnitud y Duración de la Transición: La velocidad angular promedio durante el movimiento de ponerse de pie tiende a ser mayor que durante sentarse, reflejando la naturaleza más explosiva de este movimiento. Sin embargo, las cuatro características seleccionadas capturan adecuadamente esta diferencia sin necesidad de features adicionales, ya que la magnitud de ω_{knee} y ω_{hip} inherentemente codifica esta información.

4) Robustez del Modelo con Características Compartidas: La utilización del mismo conjunto de características para sentarse y ponerse de pie demuestra la eficiencia del espacio de features seleccionado. No se requieren características específicas para cada actividad, sino que la diferenciación emerge naturalmente de los patrones en los valores de las features compartidas. Este enfoque minimalista reduce la complejidad del sistema, facilita la interpretabilidad del modelo y minimiza el riesgo de sobreajuste.

La simetría conceptual entre ambas actividades implica que un modelo entrenado para clasificar estas transiciones aprende esencialmente a reconocer dirección y magnitud de cambios articulares, habilidades que generalizan bien a otros movimientos biomecánicamente relacionados.

VI. Turning

A. Características Específicas de la Actividad

La actividad de girar representa una rotación del cuerpo sobre el eje vertical, caracterizada fundamentalmente por el cambio en la orientación del torso con respecto a la cámara. A diferencia de sentarse y ponerse de pie, que son movimientos sagitales dominados por flexión-extensión, el giro es un movimiento transversal que requiere un enfoque diferente en la extracción de características.

1) Vector de Orientación de Hombros: La característica fundamental para detectar giros es el vector de orientación de los hombros en el plano horizontal XZ. Este vector se define como:

$$\mathbf{v}_{shoulder}(t) = \mathbf{p}_{left\_shoulder}(t) - \mathbf{p}_{right\_shoulder}(t) \tag{15}$$

Para eliminar la influencia de inclinaciones verticales del torso, se proyecta este vector en el plano XZ (descartando la componente Y):

$$\mathbf{v}^{XZ}_{shoulder}(t) = [v_x(t), v_z(t)]^T \tag{16}$$

La normalización del vector garantiza que la métrica sea invariante a la distancia de la persona respecto a la cámara:

$$\hat{\mathbf{v}}^{XZ}_{shoulder}(t) = \frac{\mathbf{v}^{XZ}_{shoulder}(t)}{\|\mathbf{v}^{XZ}_{shoulder}(t)\|} \tag{17}$$

2) Cambio Angular de Rotación ($\Delta\phi$): La característica discriminante para identificar un giro es el cambio angular del vector de hombros entre frames consecutivos. Este se calcula mediante el producto punto de vectores normalizados:

$$\cos(\Delta\phi(t)) = \hat{\mathbf{v}}^{XZ}_{shoulder}(t-1) \cdot \hat{\mathbf{v}}^{XZ}_{shoulder}(t) \tag{18}$$

$$\Delta\phi(t) = \arccos\left(\text{clip}\left(\hat{\mathbf{v}}^{XZ}_{shoulder}(t-1) \cdot \hat{\mathbf{v}}^{XZ}_{shoulder}(t), -1, 1\right)\right) \times \frac{180}{\pi} \tag{19}$$

donde la función clip asegura que el argumento del arcocoseno permanezca en el dominio válido $[-1, 1]$, manejando errores numéricos.

Un giro significativo se identifica cuando $\Delta\phi(t)$ supera un threshold típico de 20 por frame. Este valor discrimina efectivamente rotaciones deliberadas del cuerpo de oscilaciones naturales o ajustes posturales menores.

3) Independencia de Otras Características Articulares: Una propiedad fundamental de la actividad de girar es su relativa independencia de las configuraciones articulares de las extremidades inferiores. Una persona puede girar estando de pie, sentada, o durante una transición. Esta independencia justifica que $\Delta\phi$ sea suficiente como característica única para esta actividad, sin necesidad de incorporar información de ángulos de cadera o rodilla.

La correlación empírica entre $\Delta\phi$ y las características articulares (θ_{knee}, θ_{hip}) es típicamente baja ($r < 0.3$), confirmando que el giro captura información complementaria y no redundante respecto a las actividades anteriores.

4) Robustez ante Oclusiones Parciales: El uso del vector de hombros como base para la detección de giros presenta ventajas robustas. Los landmarks de hombros son generalmente más estables y menos susceptibles a oclusiones que los de extremidades inferiores, particularmente en escenarios donde la persona está parcialmente detrás de objetos. Adicionalmente, la simetría bilateral permite detectar el giro incluso si uno de los hombros experimenta errores transitorios de detección, ya que el vector puede ser estimado por continuidad temporal.

VII. Resultados: Walk Forward vs Walk Backward

A. Proceso de Entrenamiento

El modelo se entrenó durante 52 epochs antes de que Early Stopping detuviera el proceso. Los mejores pesos se restauraron del epoch 37.

Curva de aprendizaje:

- **Epoch 1**: Train acc = 95.78%, Val acc = 100%
- **Epoch 15**: Val loss mínimo inicial = 0.0409
- **Epoch 37**: Mejor modelo (Val loss = 0.0138)
- **Epoch 44**: ReduceLROnPlateau activado ($\alpha = 0.0005$)
- **Epoch 52**: Early stopping activado

B. Métricas de Evaluación

TABLE I
Resultados del Modelo LSTM Bidireccional

Métrica	Train	Val	Test
Accuracy	∼100%	100%	100%
Loss	0.009	0.0138	0.0170
AUC	1.0	1.0	1.0
Precision	-	-	1.00
Recall	-	-	1.00
F1-Score	-	-	1.00

C. Matriz de Confusión

La matriz de confusión en el conjunto de test revela una clasificación perfecta:

$$\begin{bmatrix} 112 & 0 \\ 0 & 116 \end{bmatrix} \tag{20}$$

- **True Negatives (TN)**: 112 (walk_forward correctos)
- **False Positives (FP)**: 0
- **False Negatives (FN)**: 0
- **True Positives (TP)**: 116 (walk_backward correctos)

D. Classification Report

	precision	recall	f1-score	support
walk_forward	1.00	1.00	1.00	112
walk_backward	1.00	1.00	1.00	116
accuracy			1.00	228
macro avg	1.00	1.00	1.00	228
weighted avg	1.00	1.00	1.00	228

VIII. Discusión

A. Análisis del Desempeño Excepcional

El modelo alcanzó métricas perfectas (100% accuracy) en todos los conjuntos. Este resultado, aunque positivo, requiere análisis crítico:

1) Separabilidad de las Clases: Las actividades walk_forward y walk_backward presentan diferencias fundamentales que las hacen linealmente separables:

1) **Movement direction Z**: Valores completamente opuestos (positivo vs negativo)
2) **Body orientation**: Diferencia de ∼180° (frontal vs espalda)
3) **Velocity patterns**: Perfiles temporales invertidos

Esta separabilidad natural explica el alto rendimiento sin necesariamente indicar overfitting.

2) Evidencia contra Overfitting Clásico: **Argumentos a favor de la validez del modelo**:

- Val Loss ≈ Test Loss (0.0138 vs 0.0170, diferencia de 0.0032)
- Early stopping activado en epoch 52 (de 100 posibles)
- Restauración de pesos del epoch 37 (15 epochs antes del final)
- Split estratificado por video (sin data leakage entre train/test)
- Normalización correcta (solo con estadísticas de train)

3) Limitaciones del Dataset Actual: El desempeño perfecto también revela limitaciones:

1) **Dataset pequeño**: 53 videos totales
2) **Homogeneidad**: Source-1 con condiciones muy controladas
3) **Variabilidad limitada**: Mismo setup de cámara, fondo, iluminación
4) **Pocas personas**: Variabilidad fisiológica limitada

B. Impacto del Feature Engineering

1) Efectividad de la Normalización: La normalización por altura demostró ser crucial:

- Features con media extrema (> 100): Solo ángulos en grados (esperado y correcto)
- Distancias y ratios: Rangos consistentes entre [0, 3]
- Velocidades: Centradas en 0 (movimiento simétrico)

2) Feature Selection: La reducción de 50 a 39 features eliminó redundancia sin pérdida de información:

- 4 features con varianza $< 10^{-6}$: Correctamente eliminadas
- Features bilaterales: Parcialmente redundantes pero conservadas por valor discriminativo
- Features temporales: 50% eliminadas (acceleration/jerk poco informativos para estas actividades)

C. Comparación con Baselines Teóricos

Random Forest (no implementado) típicamente alcanza:

- Accuracy: 85-90% en problemas similares
- F1-Score: 0.80-0.85

Regresión Logística (lineal) esperaría:

- Accuracy: 75-80% (por separabilidad casi lineal)
- F1-Score: 0.70-0.75

El LSTM supera estos baselines, justificando su complejidad.

D. Predicciones de Generalización

El modelo actual se espera que falle en:

1) **Nuevas condiciones de captura**:
   - Ángulos de cámara diferentes (cenital, lateral extremo)
   - Iluminación variable (contraluz, sombras)
   - Oclusiones parciales
2) **Variabilidad de sujetos**:
   - Personas con movilidad reducida
   - Alturas o complexiones extremas
   - Estilos de caminar muy atípicos
3) **Velocidades extremas**:
   - Caminar muy lento (elderly)
   - Caminar muy rápido (running)

Con la integración de source-2, se espera:

- Train accuracy: 98-99%
- Val accuracy: 96-97%
- Test accuracy: 95-96%

Estas métricas serían más realistas y confiables.

IX. Conclusiones

A. Logros Principales

Este estudio ha demostrado la viabilidad de un sistema completo de reconocimiento de actividades humanas basado en pose estimation y deep learning. Los logros específicos incluyen:

1) **Pipeline robusto**: Implementación de extremo a extremo desde video crudo hasta clasificación, siguiendo mejores prácticas de ingeniería de ML.
2) **Feature engineering efectivo**: Diseño de 50 features invariantes a fisiología, con reducción exitosa a 39 features discriminativas.
3) **Modelo de alto rendimiento**: LSTM bidireccional que alcanza 100% accuracy en clasificación binaria de actividades altamente distinguibles.
4) **Metodología rigurosa**: Prevención de data leakage mediante split por video, normalización correcta y feature selection basada en evidencia.

B. Limitaciones Identificadas

1) **Dataset limitado**: 53 videos insuficientes para generalización robusta.
2) **Homogeneidad**: Condiciones de captura muy controladas reducen variabilidad.
3) **Scope reducido**: Solo 2 de 5 actividades objetivo completadas en esta iteración.
4) **Falta de validación cruzada entre datasets**: No se probó con videos de fuentes externas.

C. Trabajo Futuro

1) Corto Plazo:

1) **Técnicas de balanceo**: Si se observa desbalance, aplicar SMOTE o class weights.

2) Mediano Plazo:

1) **Data augmentation avanzado**: Transformaciones de perspectiva, simulación de oclusiones.
2) **Transfer learning**: Pre-entrenamiento en datasets públicos

3) Largo Plazo:

1) **Aplicaciones específicas**:
   - Sistema de monitoreo de caídas para adultos mayores
   - Análisis biomecánico en rehabilitación física
   - Interfaz de control gestual
2) **Despliegue en edge devices**: Implementación en dispositivos móviles o Raspberry Pi.
3) **Interpretabilidad**: Visualizaciones de atención temporal (qué frames son más importantes para la decisión).

Agradecimientos

Los autores agradecen a la Universidad ICESI y al Departamento de Computación y Sistemas Inteligentes por el apoyo en la realización de este proyecto. Especial reconocimiento al profesor del curso de Inteligencia Artificial I por la guía metodológica.

References

[1] 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.

[2] D. Tran et al., "Learning Spatiotemporal Features with 3D Convolutional Networks," *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 4489-4497, 2015.

[3] K. Simonyan and A. Zisserman, "Two-Stream Convolutional Networks for Action Recognition in Videos," *Advances in Neural Information Processing Systems*, pp. 568-576, 2014.

[4] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7291-7299, 2017.

[5] C. Lugaresi et al., "MediaPipe: A Framework for Building Perception Pipelines," *arXiv preprint arXiv:1906.08172*, 2019.

[6] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Computation*, vol. 9, no. 8, pp. 1735-1780, 1997.

[7] A. Vaswani et al., “Attention is All You Need,” *Advances in Neural Information Processing Systems*, pp. 5998-6008, 2017.
[8] MediaPipe Solutions Guide, “Pose Landmark Detection,” Google AI, https://ai.google.dev/edge/mediapipe/solutions/vision/pose_landmarker, 2024.
[9] F. Chollet et al., “Keras Documentation,” https://keras.io, 2015.
[10] A. Géron, “Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow,” 2nd ed., O’Reilly Media, 2019.