Sistema de Reconocimiento de Actividades Humanas mediante Visión por Computadora: Diseño Experimental y Consideraciones Éticas

William Joseph Verdesoto, Luis Manuel Rojas, Juan Camilo Corrales Osvath

Resumen

Se presenta el diseño experimental para un sistema de reconocimiento de actividades humanas basado en visión por computadora. El sistema debe clasificar cinco actividades cotidianas (caminar hacia la cámara, caminar de regreso, girar, sentarse y ponerse de pie) mientras realiza seguimiento de las articulaciones corporales para análisis biomecánico. El trabajo se enfoca en tres aspectos fundamentales: primero, el diseño de un protocolo de captura de video que minimice la variabilidad no controlada; segundo, la definición de una población de estudio diversa que permita entrenar modelos generalizables; y tercero, el análisis de las consideraciones éticas relacionadas con el manejo de datos biométricos. Todas las decisiones metodológicas se fundamentan en revisión de literatura académica reciente.

1. Introducción y Contexto del Problema

1.1. Planteamiento del Problema

El campo del Reconocimiento de Actividades Humanas (HAR) ha avanzado significativamente gracias al aprendizaje profundo, pero su aplicación práctica sigue limitada por la falta de datos de entrenamiento diversos y de alta calidad. Este proyecto aborda directamente ese desafío mediante el desarrollo de un sistema de anotación de video en tiempo real con un doble objetivo: por un lado, clasificar actividades clave (caminar, girar, sentarse y levantarse) y, por otro, realizar un seguimiento detallado de las articulaciones corporales para analizar parámetros biomecánicos como ángulos e inclinaciones del tronco.

Un sistema de estas características posee un gran potencial de aplicación en múltiples dominios. En el sector de la salud, podría ser fundamental para el seguimiento de la rehabilitación física de pacientes o para la detección temprana de deterioro funcional en adultos mayores. En el ámbito deportivo, facilitaría el análisis técnico de atletas para optimizar el rendimiento y prevenir lesiones, mientras que en el campo de la tecnología de asistencia, podría servir como base para el desarrollo de interfaces de control gestual para personas con discapacidad, transformando el análisis de movimiento en información valiosa y aplicable.

1.2. Naturaleza del Problema Computacional

El proyecto aborda dos problemas computacionales complementarios. El primero es un desafío de clasificación temporal, donde se busca asignar una etiqueta de actividad (ej. "sentarse") a una secuencia de fotogramas de video, lo que requiere analizar la evolución de los patrones visuales a lo largo del tiempo. El segundo es un problema de regresión, conocido como estimación de pose humana, que consiste en predecir las coordenadas espaciales continuas de las articulaciones clave en cada fotograma, una tarea compleja debido a desafíos como las oclusiones, la vestimenta y las variaciones de iluminación.

Para resolver estos desafíos, se emplearán herramientas de código abierto como MediaPipe para la estimación de pose y LabelStudio para la anotación de los datos. Sin embargo, el éxito de estos algoritmos está intrínsecamente ligado a la calidad y consistencia de los datos de entrada. Esta dependencia crítica es la que justifica y subraya la necesidad fundamental de un diseño experimental riguroso para la captura de los videos, asegurando que la información recolectada sea adecuada para entrenar modelos precisos y fiables.

1.3. Marco Metodológico: CRISP-DM

El desarrollo del sistema seguirá el marco CRISP-DM (Cross-Industry Standard Process for Data Mining), un proceso iterativo que ha emergido como el estándar de facto en proyectos de ciencia de datos. Schröer et al. (2021) realizaron una revisión

sistemática de la literatura y encontraron que CRISP-DM se utiliza en más del 60% de los proyectos de minería de datos reportados en las principales conferencias y revistas académicas. La popularidad de este marco se debe a su flexibilidad y aplicabilidad a diversos dominios, así como a su énfasis en la iteración y refinamiento continuo.
CRISP-DM divide el proceso en seis fases interconectadas. La primera fase, comprensión del negocio (o en este caso, comprensión del problema), implica definir claramente los objetivos del proyecto y traducirlos en requisitos técnicos específicos. La segunda fase, comprensión de los datos, se enfoca en la recolección inicial de datos y el análisis exploratorio para identificar características, patrones y problemas de calidad. La tercera fase, preparación de los datos, involucra la limpieza, transformación y anotación de los datos para hacerlos adecuados para el modelado. La cuarta fase, modelado, consiste en la selección, entrenamiento y ajuste de algoritmos de aprendizaje automático. La quinta fase, evaluación, valida el rendimiento del modelo contra métricas predefinidas y verifica que se cumplan los objetivos del proyecto. La sexta fase, despliegue, implementa el sistema en un entorno operacional.

2. Diseño del Protocolo de Captura de Video

2.1. Importancia del Diseño Experimental

La inconsistencia en la captura de datos (variaciones en ángulo, distancia o iluminación) degrada severamente el rendimiento de los modelos, como demostraron Vondrick et al. (2013). Por esta razón, un protocolo de captura riguroso no es solo una buena práctica, sino una necesidad metodológica fundamental.

2.2. Configuración del Entorno de Captura

2.2.1. Sistema de Iluminación

Dado que la iluminación y las sombras afectan críticamente la precisión de la estimación de pose, el proyecto utilizará las condiciones de luz ambiental existentes bajo un estricto control. Antes de cada grabación, se medirá la intensidad lumínica con una aplicación de luxómetro en un smartphone, y se ha establecido un umbral mínimo de 500 lux como requisito indispensable para proceder, garantizando así una calidad de datos consistente.

2.2.2. Fondo y Espacio Físico

Basado en estudios como el de Cormier et al. (2022), que demuestran que los fondos complejos con texturas o movimiento incrementan los falsos positivos en la detección de articulaciones, el experimento se realizará en un entorno estrictamente controlado.

Se utilizará un fondo sólido de color neutro y no reflectante, en un espacio despejado de al menos 4x4 metros. Además, se emplearán marcas en el suelo para estandarizar las posiciones iniciales y las trayectorias de movimiento, garantizando la consistencia entre todos los participantes.

2.3. Configuración de la Cámara

2.3.1. Selección del Dispositivo de Captura

Se utilizará un iPhone 11 Pro Max para la captura de video, elegido por su accesibilidad y calidad de imagen. Para garantizar la consistencia y mitigar el procesamiento automático del dispositivo, la cámara se configurará manualmente: se grabará a 1080p y 30 FPS, se activará el bloqueo de enfoque y exposición (AE/AF Lock), y se desactivarán funciones como el HDR automático y la estabilización cinemática. El formato de grabación será H.264 ("Más compatible") para asegurar un fácil procesamiento posterior.

Se configurará la cámara utilizando la aplicación nativa con resolución de 1080p a 30 FPS y se activará el bloqueo de enfoque y exposición (AE/AF Lock). Este bloqueo es crucial para impedir que los ajustes automáticos del dispositivo introduzcan variabilidad indeseada en el brillo o el foco durante y entre las grabaciones.

2.3.2. Parámetros Técnicos de Captura

La configuración técnica de la cámara se ha determinado mediante un balance entre calidad de

imagen, requisitos de almacenamiento, y compatibilidad con los algoritmos de procesamiento que se utilizarán posteriormente.

Parámetro	Configuración	Justificación
Resolución	1080p (1920 × 1080 píxeles)	Proporciona suficiente detalle espacial para que los algoritmos de estimación de pose detecten articulaciones con precisión, sin generar archivos excesivamente grandes.
Frame Rate	30 FPS (fotogramas por segundo)	Captura suficiente información temporal para representar movimientos cotidianos como caminar, sentarse y levantarse. Frecuencias menores (15-20 FPS) podrían perder fases importantes del movimiento, mientras que frecuencias mayores (60+ FPS) aumentarían innecesariamente el tamaño de archivos sin aportar información adicional relevante para estas actividades.
Altura de cámara	150 cm desde el suelo hasta el centro del lente	Corresponde aproximadamente a la altura de los ojos de una persona adulta promedio, proporcionando un ángulo de visión natural que minimiza la distorsión de perspectiva de las articulaciones
Distancia cámara-sujeto	4 metros	Permite capturar el cuerpo completo del sujeto con espacio adicional alrededor, evitando que partes del cuerpo se salgan del encuadre durante movimientos como caminar o girar. Distancias menores recortarían extremidades, mientras que distancias mayores reducirían la resolución efectiva del sujeto en la imagen.
Ángulo de cámara	Perpendicular al suelo (0° vertical) y frontal al sujeto (0° horizontal)	La vista frontal proporciona visibilidad clara de las articulaciones principales (hombros, codos, muñecas, caderas, rodillas, tobillos) sin oclusiones significativas. Ángulos laterales u oblicuos ocultarían articulaciones

Parámetro	Configuración	Justificación
		importantes o crearían ambigüedad en la profundidad.

Adicionalmente, los elementos de configuración avanzada de la cámara que modifican la toma de datos, tales como HDR y Estabilización serán desactivados; y otros como Balance de blancos, Exposición, Enfoque y Formato de Codificación serán empleados con los valores por defecto o automáticos de los dispositivos de grabación.

2.4. Protocolo de Grabación Paso a Paso

El protocolo de grabación inicia con una preparación rigurosa del equipo, que incluye la verificación de la batería y el almacenamiento del dispositivo, la limpieza del lente y la grabación de un video de prueba para validar el encuadre, el enfoque y la iluminación. A la llegada del participante, se le explica el procedimiento, se obtiene su consentimiento informado y se registran sus datos demográficos y antropométricos (edad, género, estatura y peso).

Para cada actividad, se instruye y permite practicar al participante antes de grabar. Cada toma se revisa de inmediato para asegurar la calidad; si no cumple con los criterios de visibilidad, enfoque e iluminación, se descarta y se repite. Se capturan tres repeticiones de cada actividad solicitando una ejecución a velocidad "normal", "lenta" y "rápida" para capturar la variabilidad del movimiento. Finalmente, todos los archivos se nombran sistemáticamente con la convención VIDEO[ID]_[VELOCIDAD].mp4 para una correcta organización.

3. Diseño de la Población de Estudio

El movimiento humano varía significativamente entre individuos, influenciado por factores antropométricos clave. Para este estudio, se ha seleccionado la estatura como la variable principal a controlar. Esta elección se fundamenta en que la estatura impacta directamente en la longitud de las extremidades, la cinemática del movimiento (como la longitud de la zancada) y los ángulos articulares, factores que pueden alterar la clasificación de una misma actividad por parte del modelo.

La población de estudio estará compuesta por estudiantes universitarios. La estrategia de muestreo buscará una distribución equitativa de género; sin embargo, el criterio fundamental para la selección de los participantes será priorizar la máxima variabilidad en la estatura. El objetivo es asegurar que el conjunto de datos represente un amplio rango de constituciones físicas, desde individuos de baja hasta alta estatura, para mejorar la capacidad de generalización del modelo.

3.3. Determinación del Tamaño de Muestra

La pregunta de cuántos participantes se necesitan no tiene una respuesta única, sino que depende de múltiples factores incluyendo la complejidad del problema, la variabilidad en la población, y las restricciones prácticas de tiempo y recursos. Desde una perspectiva estadística, el tamaño de muestra debería determinarse mediante un análisis de poder estadístico, que calcula cuántos participantes se necesitan para detectar un efecto de cierto tamaño con una probabilidad dada. Sin embargo, de acuerdo con indicaciones dadas por el profesor del curso, decidimos que contar con un mínimo de 15 videos es suficiente.

3.4. Distribución Demográfica de la Muestra

La composición demográfica de la muestra debe diseñarse para capturar diversidad en las variables que se identificaron como relevantes, que en el caso del experimento supone meramente la altura. Se propone incluir participantes en tres rangos (bajo, medio, alto), con 4 participantes en cada rango. Los rangos específicos son definidos a continuación: baja (150-165 cm), media (166-175 cm), alta (176-195 cm).

3.5. Variables del Experimento

Las variables de participante que se registrarán incluye solamente la estatura (medida en centímetros.

Esta variable no se puede manipular, pero se registra porque pueden afectar los patrones de movimiento y porque permitirán analizar si el modelo tiene rendimiento diferencial entre subgrupos de estatura.

Las variables de captura que se controlarán y mantendrán constantes incluyen la altura de la cámara (150 cm), la distancia cámara-sujeto (6 metros), el ángulo de la cámara (0 grados, frontal), la resolución (1080p), el frame rate (30 FPS), y la configuración de iluminación (sistema de tres puntos con intensidad de 500-700 lux). Mantener estas variables constantes es esencial para minimizar la variabilidad no controlada.

Dado el enfoque de grabación continua, las variables de actividad se registrarán mediante anotaciones temporales. Las etiquetas de clasificación corresponderán a las 5 clases de movimiento definidas: **caminar hacia la cámara, caminar de regreso, girar, sentarse y levantarse**. Para capturar una variabilidad controlada, cada participante realizará una "coreografía" completa que integra estas actividades. Dicha coreografía se grabará tres veces, instruyendo al participante a ejecutarla a una velocidad "normal", "lenta" y "rápida". El resultado será un conjunto de videos continuos que, mediante el uso de Label Studio, serán segmentados y etiquetados para identificar con precisión los instantes de inicio y fin de cada actividad, permitiendo así un análisis detallado de cómo la velocidad de ejecución afecta las características del movimiento.

Las variables dependientes principales son las métricas de rendimiento del modelo. Para la clasificación de actividades, estas incluyen la precisión global (accuracy), la precisión por clase (precision), la sensibilidad por clase (recall), el F1-Score (media armónica de precision y recall), y la matriz de confusión. Para la estimación de pose, las métricas incluyen la tasa de detección de landmarks (porcentaje de fotogramas donde se detectan todas las articulaciones), la confianza promedio de detección (score proporcionado por MediaPipe), y la suavidad temporal de las trayectorias.

Las variables confusoras que deben controlarse incluyen la vestimenta (se solicitará a los participantes usar ropa ajustada o semi-ajustada de colores sólidos, evitando patrones complejos que puedan interferir con la detección de articulaciones), el calzado (zapatos deportivos cerrados y planos, evitando tacones o sandalias), y los accesorios (se solicitará remover relojes grandes, pulseras voluminosas, o cualquier accesorio que pueda ocultar articulaciones).

4. Métricas de Evaluación

4.1. Métricas Durante la Recolección de Datos

Se monitoreará la calidad del proceso de recolección evaluando la **tasa de reclutamiento** para ajustar la captación, la **tasa de completitud** (objetivo > 90%) para validar el diseño del protocolo, y la **tasa de aceptación de videos** (> 95%) para asegurar la calidad del material. Además, se controlará la **diversidad de la muestra** para garantizar la representatividad demográfica planificada.

4.2. Métricas de Clasificación de Actividades

El rendimiento del modelo de clasificación se medirá con la **precisión (precision)**, **sensibilidad (recall)** y la **matriz de confusión**. El indicador principal será el **F1-Score**, buscando un valor promedio **superior a 0.85** para alinearse con benchmarks de referencia en el reconocimiento de actividades humanas (HAR).

4.3. Métricas de Estimación de Pose

La calidad de las coordenadas corporales se evaluará mediante la **tasa de detección** de articulaciones (objetivo > 95%), la **confianza promedio** de las predicciones (debe superar 0.7) y la **suavidad temporal** para identificar inconsistencias o ruido (“jitter”) en el movimiento.

4.4. Métricas de Equidad (Fairness)

Para garantizar la equidad, se calculará el **F1-Score por subgrupo** demográfico y se medirá la

disparidad del F1-Score entre ellos. El modelo se considerará justo si esta diferencia máxima es **menor a 0.05**; de lo contrario, se aplicarán técnicas de mitigación de sesgos.

5. Estrategias para Incrementar el Conjunto de Datos

5.1. Ampliación del Reclutamiento
La estrategia más directa es continuar el proceso de reclutamiento para duplicar el número de participantes si los resultados iniciales demuestran que se necesita más variabilidad.

5.2. Aumento de Datos (Data Augmentation)
Consiste en generar nuevas muestras sintéticas aplicando transformaciones a los videos existentes. Estas incluyen:

- Espaciales: Rotaciones, traslaciones y cambios de escala.
- De Color: Ajustes de brillo, contraste y saturación.
- Temporales: Variación de la velocidad de reproducción.

5.3. Uso de Datasets Públicos para Transfer Learning

Se propone emplear aprendizaje por transferencia (transfer learning), que consiste en pre-entrenar un modelo en un dataset público a gran escala (como Kinetics-700 o UCF101) y luego ajustarlo (fine-tuning) con los datos específicos de nuestro proyecto.

5.4. Uso de videos de compañeros de curso

Se plantea la colaboración con otros equipos del curso para compartir los videos grabados por cada grupo, creando así un dataset interno más grande y diverso de manera sencilla y accesible.

6. Análisis Exploratorio de Datos

Previo al entrenamiento de los modelos de clasificación, se realizará un análisis exploratorio de datos (EDA) exhaustivo. El objetivo de este proceso será inspeccionar y validar la calidad de los nuevos datos de video, comprender la distribución de las actividades y transformar las coordenadas 3D de los landmarks en características biomecánicas robustas que faciliten el aprendizaje del modelo.

6.1. Carga y Estructuración de los Datos

Se iniciará el proceso cargando y fusionando dos archivos JSON: uno con las coordenadas 3D de los landmarks corporales de MediaPipe y otro con las anotaciones temporales de Label Studio. Se asociará cada fotograma con su etiqueta de actividad correspondiente, descartando cualquier dato no etiquetado para asegurar que el análisis se centre exclusivamente en la información relevante.

6.2. Análisis de Calidad de Datos y Visibilidad

Una vez cargados los datos, se procederá a una inspección inicial para evaluar su integridad, verificando la ausencia de valores nulos que pudieran afectar el procesamiento posterior.
Un componente central de esta fase será el análisis del score de visibilidad proporcionado por MediaPipe. Este score es un indicador clave de la confianza en la detección de cada landmark. Para cuantificar la calidad general de la captura de datos, se realizarán las siguientes validaciones:

- Se calculará la visibilidad promedio de todos los landmarks.
- Se determinará el porcentaje de detecciones que se encuentren por debajo de un umbral de confianza predefinido (ej. 0.5) para identificar la cantidad de datos de baja calidad.
- Se analizará la visibilidad promedio por cada articulación específica, con el fin de detectar problemas sistemáticos de oclusión o seguimiento en puntos corporales clave (ej. muñecas, tobillos).

6.3. Distribución y Duración de Actividades

Para analizar la composición del dataset, se evaluará el balance de clases contando los fotogramas por actividad para detectar desequilibrios que puedan

requerir re-muestreo o ponderación durante el entrenamiento. Simultáneamente, se calculará la duración promedio y la variabilidad de los segmentos de cada actividad para asegurar que los tiempos sean coherentes con la naturaleza de los movimientos e identificar posibles ejecuciones anómalas.

6.4. Análisis de Características Biomecánicas

Para que el modelo aprenda patrones de movimiento en lugar de posiciones absolutas, se aplicará un pipeline de preprocesamiento y extracción de características.

Normalización de Coordenadas: Se establecerá un nuevo sistema de coordenadas relativo al cuerpo. Para cada fotograma, se calculará el punto central entre las caderas y se definirá como el origen (0,0,0). Todas las coordenadas de los demás landmarks se recalcularán con respecto a este nuevo origen. Este paso asegurará que las características resultantes sean invariantes a la posición y escala de la persona en el video.

Creación de Características: A partir de las coordenadas normalizadas, se generará un conjunto de características biomecánicas de alto nivel, diseñadas para capturar la dinámica del movimiento humano. Estas incluirán:

- **Ángulos Articulares:** Se calcularán los ángulos 3D para articulaciones clave como las rodillas y las caderas.
- **Inclinación del Tronco:** Se medirá el ángulo del torso con respecto a la vertical.
- **Velocidades Lineales:** Se computará la velocidad de cada landmark para capturar la dinámica del movimiento.
- **Métricas Adicionales:** Se crearán características como la distancia entre los pies, la altura relativa de la nariz (para detectar si la persona se agacha) y su distancia en el eje Z (para medir el acercamiento/alejamiento).

Análisis de Correlación: Una vez generadas las características, se calculará una matriz de correlación. El objetivo será identificar relaciones lineales fuertes entre ellas, detectar posibles redundancias (ej. entre movimientos de extremidades simétricas) y obtener una mejor comprensión de cómo interactúan las distintas variables.

6.5. Análisis de Separabilidad de Clases

Para evaluar preliminarmente si las características generadas son suficientemente discriminativas, se emplearán técnicas de reducción de dimensionalidad para visualizar la separabilidad de las clases. Se aplicará PCA para obtener una proyección lineal de los datos, lo que permitirá observar la formación de conglomerados básicos.

Adicionalmente, se utilizará t-SNE para crear una visualización no lineal que revele estructuras de clústeres más complejas, proporcionando una intuición visual más clara sobre qué tan distinguibles son las actividades. La conclusión de este análisis será un conjunto de datos limpio, validado y enriquecido, listo para la fase de entrenamiento de modelos.

7. Consideraciones Éticas y Tratamiento de Datos

En el nivel de los datos, las decisiones éticas se relacionan con cómo se obtienen, almacenan, y utilizan los datos. Este nivel es particularmente crítico para este proyecto porque involucra datos biométricos (imágenes de personas) que son inherentemente sensibles.

7.1. Consentimiento Informado y Derechos del Participante

Se implementará un proceso de **consentimiento informado** previo, expreso y documentado. A cada voluntario se le entregará un formato claro que detalla la finalidad del estudio, los procedimientos, el carácter voluntario de su participación y el tratamiento de sus datos, que son considerados sensibles por su naturaleza biométrica. Se garantizarán los derechos de los titulares a conocer, actualizar, rectificar y solicitar la supresión de sus datos en cualquier momento, estableciendo un canal

de comunicación directo para atender dichas solicitudes (1143878894@u.icesi.edu.co).

7.2. Anonimización y Seguridad de los Datos

Para proteger la identidad de los participantes, se implementará un protocolo de **seudonimización**, asignando un código alfanumérico único a cada voluntario para nombrar los archivos de video. La información personal identificable se almacenará de forma separada y cifrada. Se aplicará un difuminado de rostro a cualquier material visual utilizado en divulgaciones académicas. Los datos se almacenarán en repositorios privados y cifrados con acceso restringido. La política de retención establece que los datos serán conservados hasta 12 meses después de la finalización del curso, tras lo cual se procederá a su supresión segura e irreversible.

7.3. Mitigación de Sesgos y Uso Responsable

Conscientes del riesgo de sesgos algorítmicos, el diseño del reclutamiento busca activamente una muestra diversa en términos de género, complexión y altura. Se analizará el rendimiento del modelo en diferentes subgrupos para detectar disparidades y se reportarán de manera transparente las limitaciones del sistema. Se reitera que la tecnología desarrollada se utilizará únicamente con fines educativos y de investigación, y no será desplegada en aplicaciones que puedan comprometer la privacidad o tomar decisiones críticas sobre individuos.

Referencias

1. Universidad ICESI. (2025). Lineamientos para el proyecto final - Algoritmos y Programación III. Documento del curso.

2. Schröer, C., Kruse, F., & Gómez, J. M. (2021). A Systematic Literature Review on Applying CRISP-DM Process Model. Procedia Computer Science, 181, 526-534. https://www.sciencedirect.com/science/article/pii/S1877050921002416
3. Vondrick, C., Patterson, D., & Ramanan, D. (2013). Efficiently Scaling up Crowdsourced Video Annotation. International Journal of Computer Vision, 101(1), 184-204. https://link.springer.com/article/10.1007/s11263-012-0564-1
4. Cormier, M., Wiliem, A., Lovell, B. C., & Mansur, A. (2022). Where are we with human pose estimation in real-world surveillance? IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2767-2776. https://arxiv.org/abs/2110.03560
5. Kay, W., Carreira, J., Simonyan, K., et al. (2017). The Kinetics Human Action Video Dataset. arXiv preprint arXiv:1705.06950. https://arxiv.org/abs/1705.06950
6. Soomro, K., Zamir, A. R., & Shah, M. (2012). UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild. arXiv preprint arXiv:1212.0402. https://arxiv.org/abs/1212.0402