Universidad Zaragoza
1542

Trabajo Fin de Grado

Generación de Base de Datos mediante la Estimación de Puntos Clave Corporales y Detección de Pose por Algoritmos de Inteligencia Artificial

Database Generation through Body Keypoint Estimation and Pose Detection Using Artificial Inteligence Algorithms

Autor

David Polo Llimós

Directores

Pablo Pérez Lázaro & Rocío Aznar Gimeno

Escuela Universitaria Politécnica La Almunia

Junio 2025

Página intencionadamente en blanco.

Escuela Universitaria Politécnica - La Almunia
Centro adscrito
Universidad Zaragoza

ESCUELA UNIVERSITARIA POLITÉCNICA
DE LA ALMUNIA DE DOÑA GODINA (ZARAGOZA)

MEMORIA

Generación de Base de Datos mediante la Estimación de Puntos Clave Corporales y Detección de Pose por Algoritmos de Inteligencia Artificial

Database Generation through Body Keypoint Estimation and Pose Detection Using Artificial Inteligence Algorithms

625.25.69

Autor: David Polo Llimós

Director: Pablo Pérez Lázaro & Rocío Aznar Gimeno

Fecha: Junio 2025

Página intencionadamente en blanco.

INDICE DE CONTENIDO BREVE

INDICE DE CONTENIDO

INDICE DE ILUSTRACIONES

INDICE DE TABLAS

1. RESUMEN

Este trabajo explora el desarrollo de un conjunto de datos para la estimación de la pose humana en 3D, combinando grabaciones de vídeo con datos de sensores inerciales (IMUs) del traje Rokoko Smartsuit Pro II. A pesar de los avances en la estimación de pose en 3D, la mayoría de los conjuntos de datos públicos se basan en sistemas ópticos costosos y en entornos controlados, lo que limita su accesibilidad y aplicabilidad en entornos reales. El objetivo de este trabajo es la construcción de un conjunto de datos que combina datos de IMUs y vídeo, capturando 14 acciones cotidianas realizadas por 8 sujetos en entornos exteriores. El conjunto está estructurado para ser aplicado a modelos de aprendizaje profundo para la estimación de pose humana 3D. Para garantizar la calidad de los datos, se diseñó una estrategia de grabación multicámara sincronizada, junto con procedimientos específicos para alinear temporalmente las secuencias de vídeo y los registros inerciales. Los modelos del estado del arte MediaPipe, MHFormer y MotionBERT fueron evaluados en el conjunto de datos desarrollado usando la métrica PA-MPJPE. En esta evaluación se observaron comportamientos anómalos de los modelos en secuencias con oclusiones o posturas complejas. Este proyecto aporta un conjunto de datos valioso y reproducible que impulsa la estimación de pose en 3D en escenarios reales y apoya la investigación en salud, biomecánica e interacción persona-computadora.

1.1. PALABRAS CLAVE

- Estimación de pose humana 3D
- Sensores inerciales
- Captura de movimiento
- Creación de dataset
- Visión por computador

2. ABSTRACT

This work explores the development of a dataset for 3D human pose estimation, combining video recordings with inertial sensor data (IMUs) from the Rokoko Smartsuit Pro II. Despite recent advances in 3D pose estimation, most public datasets rely on expensive optical systems and controlled environments, which limits their accessibility and applicability in real-world scenarios. The objective of this work is to build a dataset that integrates IMU and video data, capturing 14 everyday actions performed by 8 subjects in outdoor settings. The dataset is structured to support the application of deep learning models for 3D human pose estimation. To ensure data quality, a synchronized multi-camera recording strategy was designed, along with specific procedures to temporally align video sequences and inertial recordings. State-of-the-art models—MediaPipe, MHFormer, and MotionBERT—were evaluated using the developed dataset and the PA-MPJPE metric. During the evaluation, anomalous behaviors were observed in sequences involving occlusions or complex postures. This project contributes a valuable and reproducible dataset that advances real-world 3D pose estimation and supports research in healthcare, biomechanics, and human–computer interaction.

2.1. KEY WORDS

- 3D Human Pose Estimation
- Inertial Measurement Units (IMUs)
- Motion Capture
- Dataset creation
- Computer Vision.

3. INTRODUCCIÓN Y MOTIVACIÓN

La estimación de la pose humana en 3D (3D-HPE, por sus siglas en inglés) es un campo en constante evolución dentro de la visión por computador, con aplicaciones en áreas como la biomecánica, la animación digital y la realidad aumentada (Xu et al., 2022). En biomecánica, por ejemplo, esta tecnología permite analizar la cinemática humana en deportes y entornos clínicos, facilitando la detección temprana de patologías neurodegenerativas y la prevención de lesiones (Aznar-Gimeno et al., 2024). Con el desarrollo de modelos basados en inteligencia artificial, como MediaPipe (Bazarevsky et al., 2020), MHFormer (Li, 2021/2025), MotionBERT (Zhu et al., 2023) y VideoPose3D (Pavllo et al., 2019), la capacidad de inferir puntos clave del cuerpo en 3D ha avanzado considerablemente, abordando desafíos como la ambigüedad en la profundidad y la oclusión.

Los enfoques tradicionales de captura de movimiento han dependido de datos recogidos con marcadores ópticos, como los utilizados por Vicon (*Award Winning Motion Capture Systems*, s. f.) y OptiTrack (*Motion Capture Systems*, s. f.). Estos sistemas emplean cámaras infrarrojas y marcadores reflectantes colocados en puntos específicos del cuerpo para reconstruir con precisión la pose en 3D. Aunque ofrecen gran exactitud y son considerados el estándar de referencia en entornos controlados, requieren una infraestructura costosa, un espacio calibrado y personal especializado para su operación.

Frente a estas limitaciones, han surgido soluciones más accesibles basadas en sensores inerciales (IMUs, por sus siglas en inglés), que permiten registrar los movimientos corporales mediante la combinación de acelerómetros, giróscopos y magnetómetros. Sistemas como el Rokoko Smartsuit Pro II (*Smartsuit Pro II - Quality body motion capture in one simple mobile mocap suit*, s. f.), el MVN Awinda de Xsens (*Xsens MVN Animate - Motion Capture Software for Professionals*, s. f.) y el Perception Neuron 3 (*Perception Neuron Series | Noitom Motion Capture Systems*, s. f.) han demostrado ser alternativas viables a los sistemas ópticos tradicionales. Estos permiten la captura del movimiento sin necesidad de cámaras externas ni condiciones de iluminación controladas, lo que facilita su uso en entornos más flexibles y naturales, incluyendo exteriores o espacios reducidos. Esta diferencia entre ambos sistemas puede apreciarse visualmente en la Figura 1, donde se comparan sus configuraciones y nivel de complejidad.

Figura 1. Comparativa entre el sistema de captura de movimiento óptico Vicon (izquierda) y el sistema basado en IMUs Rokoko Smartsuit Pro II (derecha)

Para llevar a cabo la estimación de pose humana en 3D mediante técnicas de aprendizaje automático, es imprescindible contar con conjuntos de datos (datasets) adecuados que contengan información espacial y temporal del movimiento humano. La disponibilidad y calidad de estos datos resultan fundamentales tanto para el entrenamiento como para la evaluación de los modelos.

En este sentido, existen varios conjuntos de datos utilizados en la comunidad científica. COCO (Lin et al., 2015) es ampliamente utilizado para la detección de personas y keypoints en 2D, mientras que HumanEva (Sigal et al., 2010) y Human3.6M (Ionescu et al., 2014) han sido fundamentales para la estimación de pose en 3D.

El campo de la estimación de pose en 3D sigue evolucionando con el desarrollo de arquitecturas de aprendizaje profundo como las redes neuronales convolucionales (CNNs), las redes de grafos convolucionales (GCNs) y los transformers, que han mejorado la precisión y la eficiencia de los modelos.

Dado el creciente interés en soluciones accesibles y eficientes para la estimación de pose en 3D, el desarrollo de nuevas bases de datos que integren datos de sensores inerciales con algoritmos de inteligencia artificial constituye una línea de investigación clave para la mejora y aplicabilidad de estas tecnologías.

3.1. MOTIVACIÓN

Este Trabajo de Fin de Grado (TFG) surge de mi interés por aplicar los conocimientos adquiridos a lo largo del grado de Ingeniería de Datos en un proyecto práctico de procesamiento y análisis de datos. La oportunidad de trabajar y desarrollar un conjunto de datos real y desarrollar un sistema de estimación de pose humana en 3D mediante inteligencia artificial me permitirá consolidar y profundizar conocimientos sobre el procesamiento y uso de datos.

Además, este proyecto representa una excelente oportunidad para profundizar en áreas como la visión por computador y el procesamiento

de imágenes, campos que han despertado mi interés, aunque no han sido explorados en profundidad durante la carrera.

Por otro lado, el uso de IMUs en este estudio permitirá explorar tecnologías innovadoras aplicadas a la captura del movimiento humano, con el objetivo de contribuir tanto al desarrollo de nuevas bases de datos como a la mejora en la precisión de los modelos de estimación de pose. Cabe destacar que la disponibilidad de conjuntos de datos basados en IMUs es actualmente bastante limitada, lo que refuerza la necesidad de generar nuevas bases de datos centradas específicamente en esta tecnología.

Con este TFG, aspiro a reforzar mis conocimientos técnicos y contribuir a un área de investigación en expansión, con aplicaciones relevantes en ámbitos como la salud, donde la estimación precisa de la pose en 3D puede mejorar procesos de rehabilitación, monitoreo y tratamiento de pacientes.

3.2. DESCRIPCIÓN DEL PROBLEMA

La estimación de la pose humana en 3D es un problema complejo en visión por computador que consiste en predecir las coordenadas tridimensionales de las articulaciones (keypoints) del cuerpo humano a partir de imágenes. Este problema presenta múltiples desafíos debido a la ambigüedad de profundidad en imágenes monoculares, la oclusión parcial o total de ciertas partes del cuerpo y la variabilidad de las condiciones de iluminación, ropa y entornos.

Tradicionalmente, la obtención de datos de referencia para este tipo de tareas se ha realizado mediante sistemas ópticos con marcadores reflectantes, como los ofrecidos por Vicon u OptiTrack. Aunque estos sistemas proporcionan una alta precisión y fiabilidad, presentan limitaciones importantes: requieren un entorno controlado, calibración cuidadosa, múltiples cámaras sincronizadas y una infraestructura costosa, lo que restringe su uso a laboratorios especializados.

Como alternativa, los IMUs han surgido como una opción más accesible y flexible. Estos dispositivos miden el movimiento y la orientación de un cuerpo en el espacio mediante acelerómetros, giroscopios y, en algunos casos, magnetómetros. Aunque tienen una precisión ligeramente inferior, no dependen de cámaras especiales ni de un espacio controlado. Esto permite capturar movimiento en entornos reales y no restringidos, haciendo posible registrar actividades que no podrían realizarse fácilmente en laboratorio, como jugar al fútbol o correr en exteriores. En este TFG se ha trabajado con el traje Rokoko Smartsuit Pro II (*Smartsuit Pro II - Quality body motion capture in one simple mobile mocap suit*, s. f.), un sistema portátil compuesto por 19 IMUs

distribuidos por todo el cuerpo, que captura los movimientos en tiempo real mediante una red inalámbrica y un software propio.

El uso de este traje ha implicado varios retos técnicos. En primer lugar, aprender a calibrar y hacer funcionar el sistema. Por otro lado, el uso de un dispositivo Rokoko Coil Pro, cuya función es mejorar la precisión de la captura de datos. Este dispositivo permite conocer la posición del sujeto respecto al propio dispositivo. Por último, la sincronización de la captura de datos del traje con las imágenes externas de las cámaras, necesarias para el entrenamiento del modelo.

Hasta donde se sabe, apenas no existen conjuntos de datos públicos que combinen capturas de movimiento mediante sensores inerciales con anotaciones precisas de keypoints en 3D sincronizadas con imágenes. Este TFG busca abordar esta carencia mediante la creación de un conjunto de datos específico utilizando el traje Rokoko Smartsuit Pro II. Una vez generado, se evaluará la capacidad de distintos modelos preexistentes de aprendizaje profundo para estimar poses humanas en 3D a partir de imágenes, utilizando nuestro conjunto de datos como referencia para la validación.

3.3. OBJETIVOS Y METODOLOGÍA

El objetivo principal de este Trabajo de Fin de Grado es la recopilación y construcción de un dataset propio que integre datos de un traje de captura de movimiento basado en IMUs junto con grabaciones de vídeo, utilizando el traje Rokoko Smartsuit Pro II. A partir de esta base de datos, se pueden evaluar y validar modelos de deep learning para la estimación de pose en 3D a partir de imágenes 2D.

Para llevar a cabo este objetivo, se siguió la metodología iterativa CRISP-DM (Cross Industry Standard Process for Data Mining), un enfoque ampliamente utilizado para proyectos de datos. Esta metodología proporciona un marco estructurado y flexible que facilita la gestión de proyectos de análisis de datos en distintas fases, asegurando un desarrollo organizado y reproducible:

- Business Understanding: Se define el objetivo del proyecto desde una perspectiva contextual y de negocio.
- Data Understanding: Se recopilan los datos iniciales y se exploran para familiarizarse con su contenido.
- Data Preparation: Se transforman y limpian los datos para que estén listos para el modelado.
- Modeling: Se seleccionan y entrenan los algoritmos de aprendizaje automático adecuados.
- Evaluation: Se evalúa el rendimiento del modelo según métricas relevantes y se revisa si cumple los objetivos definidos en la primera fase.

- Deployment: Se pone en marcha el modelo para su uso real.

En la Figura 2 se muestra la relación entre las distintas fases de dicha metodología:

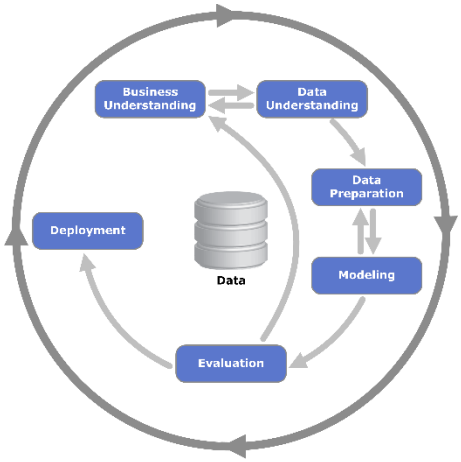

Figura 2. Esquema de los pasos de la metodología CRISP-DM

Siguiendo esta metodología, para alcanzar los objetivos planteados, a continuación, se detallan los retos que aborda este trabajo. Al final de esta sección, se muestra el diagrama de Gantt por semanas estimado asociado a esta planificación (Figura 3), el cuál fue cumplido posteriormente durante la realización del trabajo:

- Revisión del Estado del Arte: Se analiza la literatura actual sobre estimación de pose humana en 3D, abordando tanto datasets de referencia como modelos de deep learning, con especial atención a sus estructuras, anotaciones y formatos de representación de keypoints.
- Captura y preparación de los datos: Se configuran e integran los dispositivos necesarios para la captura de datos, incluyendo el traje Rokoko Smartsuit Pro II, cámaras y el sistema de geoposicionamiento Coil Pro. Esta fase comprende la puesta a punto del sistema, pruebas de calibración y extracción inicial de imágenes y keypoints.
- Procesamiento y sincronización de los datos: Se lleva a cabo la calibración de cámaras mediante checkerboard, se sincronizan los datos de vídeo y sensores, y se proyectan los keypoints sobre las imágenes. También se incluyen grabaciones multicámara para mejorar la robustez del dataset.
- Construcción del conjunto de datos: Se definen los criterios de grabación, incluyendo las condiciones del entorno, la selección de participantes, el tipo de movimientos a registrar y la configuración de cámaras, con el objetivo de obtener un conjunto de datos variado y representativo.
- Evaluación del dataset y análisis de modelos: Se valida la calidad del conjunto de datos utilizando modelos preentrenados de estimación de pose, y se comparan los resultados obtenidos con los

de otros datasets de referencia empleando métricas como el MPJPE (Mean Per Joint Position Error).

La estructura general de este trabajo ha sido diseñada siguiendo una secuencia lógica y metodológica alineada con el marco CRISP-DM. Cada capítulo aborda un reto específico dentro del proceso de minería de datos, desde la comprensión del problema y la preparación de los datos, hasta la evaluación de distintos enfoques de modelado y la interpretación de los resultados obtenidos. Esta organización garantiza una progresión coherente y completa que facilita la comprensión del lector sobre los pasos seguidos para alcanzar los objetivos del proyecto.

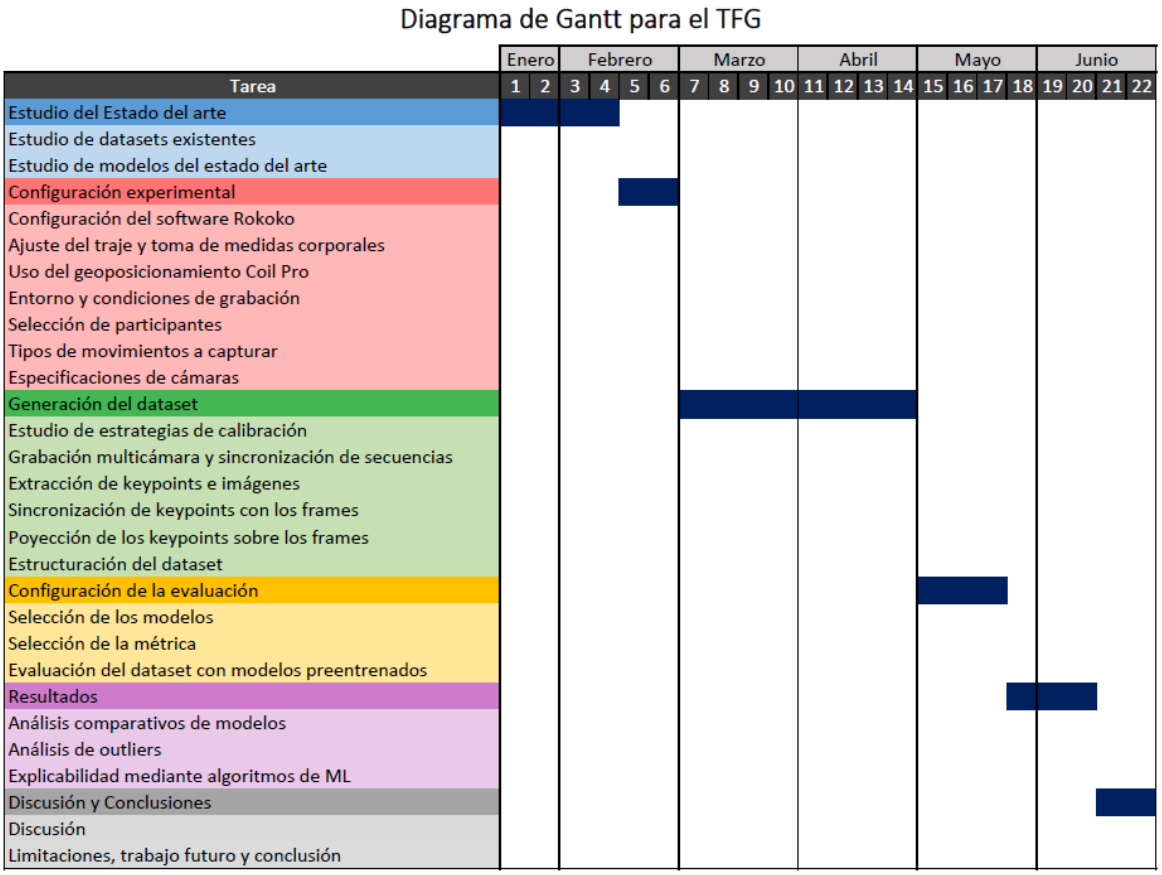

Figura 3. Diagrama de Gantt para la estructura del TFG

4. ESTADO DEL ARTE

La estimación de la pose humana en 3D es un campo en constante evolución dentro de la visión por computador. A lo largo de los años, se han desarrollado diferentes enfoques para abordar este problema. En este contexto, el presente apartado se centra en revisar los principales avances tanto en los conjuntos de datos utilizados para la estimación de pose en 3D como en las arquitecturas de modelos que han marcado el estado del arte. Se analizan los formatos, esquemas de anotación y limitaciones de los datasets más relevantes, así como las principales tendencias en los enfoques basados en aprendizaje profundo. Esta revisión proporciona el marco teórico y contextual necesario para entender contexto del trabajo desarrollado en este TFG.

4.1. DATASETS PARA LA ESTIMACIÓN DE POSE EN 3D

Los modelos de 3D-HPE dependen en gran medida de la disponibilidad de conjuntos de datos que incluyan anotaciones precisas de las articulaciones del cuerpo humano en coordenadas tridimensionales. Estos datasets son esenciales tanto para el entrenamiento como para la evaluación de los modelos. Existen principalmente dos enfoques para la obtención de estos datos: por un lado, los basados en sistemas ópticos de captura de movimiento (MoCAP), que utilizan múltiples cámaras y marcadores en entornos controlados para registrar el movimiento con alta precisión; y por otro, los que emplean IMUs, que permiten capturar movimiento de forma más flexible en escenarios no estructurados, aunque con una precisión generalmente inferior.

4.1.1. Datasets basados en captura de movimiento

Los datasets de captura de movimiento tradicional utilizan sistemas ópticos que requieren múltiples cámaras y marcadores reflectantes adheridos al cuerpo humano. Estos sistemas ofrecen una alta precisión en las anotaciones, siendo la resolución espacial del orden de 0.1 a 2 mm, lo que los convierte en la referencia estándar para validación de modelos. Entre los más destacados se encuentran:

- Human3.6M (Ionescu et al., 2014): Es el conjunto de datos más grande y uno de los más utilizados para la estimación de la pose humana 3D. Contiene grabaciones de once actores realizando quince actividades desde 4 perspectivas distintas. Sin embargo, el grupo de investigación que lo desarrolló ha dejado de conceder acceso a los datos, lo que ha dificultado su uso en investigaciones recientes.

- HumanEva (Sigal et al., 2010): Consta de dos conjuntos de datos, uno más grande y otro con un hardware más avanzado. En total incluye a 6 sujetos aunque de nuevo no es completamente abierto al público.
- MPI-INF-3DHP (Mehta et al., 2017): Incluye sujetos en entornos indoor y outdoor con anotaciones en 3D obtenidas a partir de un sistema de captura con y sin marcadores. Este conjunto de datos sí que es abierto al público. Sin embargo, presenta algunas limitaciones como que los entronos exteriores están parcialmente controlados.

En general, estos datasets son altamente precisos, pero tienen como principales limitaciones el alto coste de los sistemas de captura y su escasa representatividad de entornos reales y no controlados. Además, muchos de ellos presentan baja diversidad demográfica y de estilos de movimiento. Otra limitación importante es que la mayoría no son abiertos, lo que restringe su disponibilidad para la comunidad investigadora.

4.1.2. Datasets basados en sensores inerciales

Los sensores inerciales (IMUs) han emergido como una alternativa a los sistemas ópticos, al permitir capturar datos de movimiento sin necesidad de cámaras ni entornos controlados. Estas ventajas hacen que las IMUs sean especialmente útiles en contextos donde la movilidad y la facilidad de uso son importantes, como deportes o rehabilitación. Sin embargo, la cantidad de datasets públicos basados exclusivamente en IMUs sigue siendo limitada, y la mayoría combina datos inerciales con otros métodos de captura. Algunos ejemplos incluyen:

- Total Capture (Trumble et al., 2018): Conjunto de datos que combina IMUs con sistemas de captura ópticos. Incluye grabaciones de 5 sujetos con un total de 8 cámaras. Sin embargo, no es completamente abierto al público y al depender de sistemas ópticos, comparte algunas de sus desventajas, como el alto costo y la necesidad de un entorno controlado para la captura de movimiento.
- 3DPW (Von Marcard et al., 2018): Incluye imágenes en interiores y exteriores junto con datos de IMUs para mejorar la estimación de pose en entornos naturales. Aunque es un dataset muy variado, se centra en acciones cotidianas y simples, por lo que no incluye movimientos complejos o deportivos.

Aunque los datasets basados en IMUs son más flexibles, presentan algunas limitaciones importantes. En primer lugar, su precisión es inferior a la de los sistemas ópticos, situándose en el rango de 1 a 5 mm. En segundo lugar, los sensores inerciales pueden sufrir de drift, un error acumulativo que aparece cuando los sensores pierden su referencia

espacial con el tiempo, lo que afecta negativamente a la precisión en largas secuencias. Todo esto plantea un reto a abordar en cuanto a la investigación y desarrollo de datasets basados en IMUs.

4.2. MODELOS DE ESTIMACIÓN DE POSE EN 3D

El desarrollo de modelos de estimación de pose en 3D ha evolucionado desde enfoques tradicionales hasta arquitecturas avanzadas de deep learning. Estos modelos pueden clasificarse en función de la estrategia utilizada para inferir la pose: aquellos que convierten poses 2D en 3D (lifting) y los que estiman directamente la pose en 3D a partir de imágenes o secuencias de video.

4.2.1. Métodos basados en lifting de 2D a 3D

Los enfoques de lifting consisten en estimar primero los keypoints en 2D sobre una imagen o secuencia de vídeo, y a continuación, aplicar un modelo que reconstruya la estructura del cuerpo humano en tres dimensiones a partir de esas coordenadas 2D.

Este enfoque permite aprovechar grandes conjuntos de datos anotados en 2D (Lin et al., 2015), que son mucho más abundantes que los datasets con anotaciones precisas en 3D. Además, es un método que reduce la complejidad del problema, dividiéndolo en dos etapas: detección y reconstrucción. Sin embargo, también conlleva desafíos, como la ambigüedad de profundidad, ya que múltiples poses en 3D pueden proyectarse desde una misma pose 2D.

Algunos modelos representativos incluyen:

- VideoPose3D (Pavllo et al., 2019): Emplea redes neuronales convolucionales y recurrentes para estimar poses 3D a partir de secuencias de video 2D.
- MHFormer (Li, 2021/2025): Utiliza transformers para modelar relaciones espaciales y temporales en la reconstrucción de poses 3D a partir de estimaciones 2D. Su arquitectura modular permite modelar relaciones espaciales y temporales complejas, logrando una alta precisión especialmente en casos de oclusión o poses poco frecuentes.
- MotionBERT (Zhu et al., 2023): Modelo basado en Transformers diseñado para estimación de pose 3D a partir de videos en 2D. Destaca por su capacidad de transferir las representaciones aprendidas a tareas adicionales como reconocimiento de acciones o reconstrucción de mallas 3D.

4.2.2. Métodos basados en detección directa 3D

Los métodos de detección directa en 3D abordan la estimación de la pose humana prediciendo directamente las coordenadas tridimensionales de las articulaciones a partir de imágenes o secuencias de vídeo, sin necesidad de una etapa intermedia de predicción en 2D. Esta aproximación presenta algunas ventajas como una mayor coherencia espacial especialmente en casos de oclusión o poses complejas. Ejemplos de modelos incluyen:

- MediaPipe (Bazarevsky et al., 2020): Una solución en tiempo real desarrollada por Google para la estimación de pose en 3D basada en redes neuronales ligeras. Sin embargo requiere de licencia para su uso.
- VoxelPose (Ye et al., 2022): Este modelo propone una arquitectura híbrida que combina información de múltiples cámaras para representar el cuerpo humano en un espacio volumétrico tridimensional. Aunque se clasifica como un método de detección directa en 3D, requiere información detallada sobre los parámetros intrínsecos y extrínsecos de las cámaras para proyectar las características en un espacio común.

4.3. MÉTRICAS DE EVALUACIÓN DE MODELOS

La evaluación de modelos de estimación de pose en 3D requiere el uso de métricas específicas que consideren las particularidades del problema. A diferencia de otras tareas en visión por computador, no todos los modelos predicen las poses humanas en la misma escala, orientación o sistema de coordenadas. Algunos outputs pueden estar en un sistema relativo al sujeto, otros en coordenadas absolutas, y las escalas pueden variar en función del dataset o del modelo. Por esta razón, resulta necesario utilizar métricas que compensen esas diferencias, como aquellas que alinean las predicciones con la referencia antes de calcular el error, además de métricas tradicionales de error medio. Entre las métricas utilizadas se encuentran (Neupane et al., 2024):

- MPJPE (Mean Per Joint Position Error): Error promedio en la predicción de la ubicación de cada articulación. En otras palabras, promedio de las distancias euclidianas entre la posición predicha y la real de las articulaciones.
- PA-MPJPE (Procrustes Aligned MPJPE): Es una variante del MPJPE que tiene en cuenta los posibles errores en la escala y rotación al comparar las poses.
- PCK (Percentage of Correct Keypoints): Evalúa la precisión de la predicción de articulaciones clave en función de una tolerancia establecida (suele establecerse en 150 mm en muchos estudios). Esta métrica se utiliza especialmente en casos donde las posiciones de las articulaciones deben coincidir con las ubicaciones verdaderas dentro de una cierta tolerancia.

4.4. DESAFÍOS ACTUALES Y TENDENCIAS FUTURAS

Pese a los avances recientes, la 3D-HPE sigue enfrentando desafíos importantes:

- Ambigüedad en la profundidad: Cuando se utiliza una única cámara RGB convencional para capturar el movimiento, la imagen obtenida no contiene información directa sobre la distancia de cada parte del cuerpo respecto a la cámara. A diferencia de las cámaras estéreo o sensores de profundidad, estas cámaras no permiten determinar con precisión la profundidad de los puntos. Esto provoca que diferentes poses en 3D puedan proyectarse en una misma imagen 2D. Por ejemplo, una persona con los brazos estirados hacia adelante puede parecer similar, en la imagen, a otra con los brazos más bajos y pegados al torso, si se observa desde el frente.
- Oclusión y ruido: En muchas escenas, ciertas partes del cuerpo pueden quedar ocultas, ya sea por otros objetos o por el propio cuerpo. Por ejemplo, en una pose lateral, una pierna o un brazo pueden bloquear visualmente la vista del otro lado del cuerpo, dificultando su detección por parte del modelo. Además, factores como sombras intensas, fondos complejos, o desenfoques introducen ruido en la imagen, afectando negativamente la precisión de los keypoints detectados.
- Escasez de datasets con IMUs: A diferencia de los sistemas ópticos, existen pocas bases de datos con datos inerciales sincronizados con imágenes RGB.

En conclusión, el campo de la estimación de pose en 3D ha evolucionado significativamente gracias a los avances en deep learning y la disponibilidad de datasets. Sin embargo, persisten desafíos que dificultan la generalización de los modelos en escenarios reales. Este TFG se posiciona dentro de esta línea de investigación, proponiendo la creación de un dataset que integre datos de sensores inerciales y vídeo, lo que permitirá explorar nuevas estrategias para mejorar la estimación de pose en 3D.

5. DESARROLLO

En este capítulo se detallan los métodos y procedimientos implementados a lo largo del proyecto. El objetivo es describir de forma clara y estructurada todas las fases del trabajo, desde la preparación del entorno experimental hasta la evaluación de distintos modelos utilizando el conjunto de datos generado.

En primer lugar, se presentan los recursos técnicos empleados, incluyendo el sistema Rokoko Smartsuit Pro II y el geoposicionamiento Coil Pro, así como los criterios seguidos para la selección del entorno de grabación, los participantes y los tipos de movimientos a registrar. También se describen las características de las cámaras utilizadas.

A continuación, se expone el proceso completo de calibración de cámaras, sincronización de dispositivos, extracción de keypoints y su proyección sobre los fotogramas correspondientes. Además, se detalla cómo se organiza y estructura el conjunto de datos final para su posterior utilización.

Finalmente, se explica la metodología empleada para evaluar distintos modelos de estimación de pose 3D preentrenados utilizando el conjunto de datos generado. También se describen las métricas aplicadas para comparar el rendimiento de los modelos.

5.1. CONFIGURACIÓN EXPERIMENTAL

Siguiendo las primeras fases de la metodología CRISP-DM, esta etapa se centra en la preparación del entorno de trabajo y la definición de las condiciones necesarias para la captura de datos. Se distingue entre la configuración técnica del sistema y la selección de los elementos que definirán las características del conjunto de datos.

En cuanto a la puesta en marcha del equipo, se realizó la configuración del software Rokoko Studio para gestionar la captura de movimiento mediante el traje Smartsuit Pro II. Se ajustó físicamente el traje a cada participante mediante la toma de medidas corporales, garantizando una correcta colocación de las IMUs. Además, se incorporó el sistema Coil Pro como referencia espacial externa (un dispositivo que genera un campo electromagnético para proporcionar una estimación más precisa de la posición absoluta en el espacio), lo que permitió mejorar la precisión en la localización absoluta del sujeto durante las sesiones.

Por otro lado, se definieron las características del conjunto de datos. Se seleccionó el entorno de grabación en el que se iba a llevar a cabo y se establecieron las condiciones técnicas de las cámaras. Los participantes fueron elegidos para aportar cierta diversidad morfológica,

y se diseñó un repertorio de movimientos que incluyera tanto acciones básicas como gestos amplios, con el objetivo de generar un conjunto de datos variado y representativo.

5.1.1. Configuración del software Rokoko

La primera fase del proceso consistió en la configuración completa del sistema Rokoko Smartsuit Pro II, asegurando la instalación del software Rokoko Studio, la conexión de los dispositivos físicos y la visualización en tiempo real de los datos capturados.

Como se ha introducido anteriormente, el Smartsuit Pro II es un sistema de captura de movimiento basado en sensores inerciales distribuidos estratégicamente por todo el cuerpo del usuario. Cada sensor (una unidad IMU) recoge información sobre la aceleración lineal y la velocidad angular de la parte del cuerpo donde está colocado. A partir de estas medidas, el software realiza una integración numérica para calcular la orientación de cada segmento corporal y, mediante modelos biomecánicos, estima la posición relativa de las articulaciones. Como los sensores no ofrecen directamente coordenadas espaciales absolutas, el sistema infiere la pose completa mediante una reconstrucción jerárquica, combinando las transformaciones locales de cada segmento a lo largo de la cadena cinemática del cuerpo.

El sistema incluye una unidad central (Hub), una batería recargable y un conjunto de 19 sensores IMU, todos integrados en un traje de tela flexible. Como se puede observar en la Figura 4, los datos recogidos se transmiten de forma inalámbrica al software Rokoko Studio, donde pueden visualizarse, registrarse y exportarse para su posterior análisis.

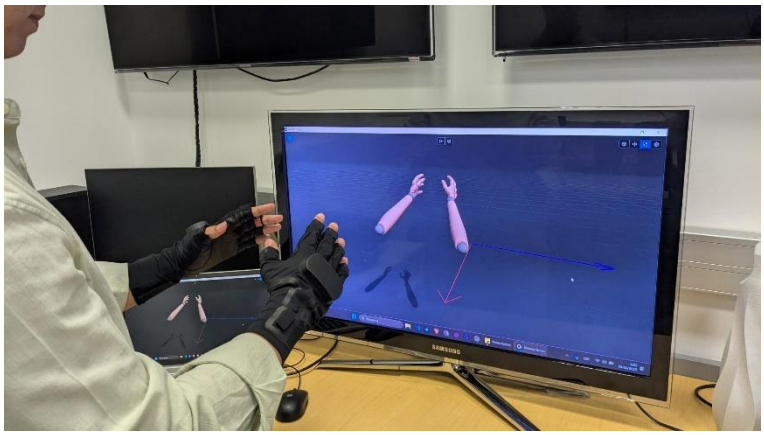

Figura 4. Visualización de las manos en el software Rokoko Studio a través de los Smart Gloves

A modo ilustrativo, en la Figura 5 se muestran tres elementos representativos del sistema Rokoko Smartsuit Pro II: una vista general del traje con los sensores integrados, una imagen del usuario portando el traje durante la fase experimental, y una visualización tridimensional

de los keypoints generados por el software Rokoko Studio a partir de los datos capturados.

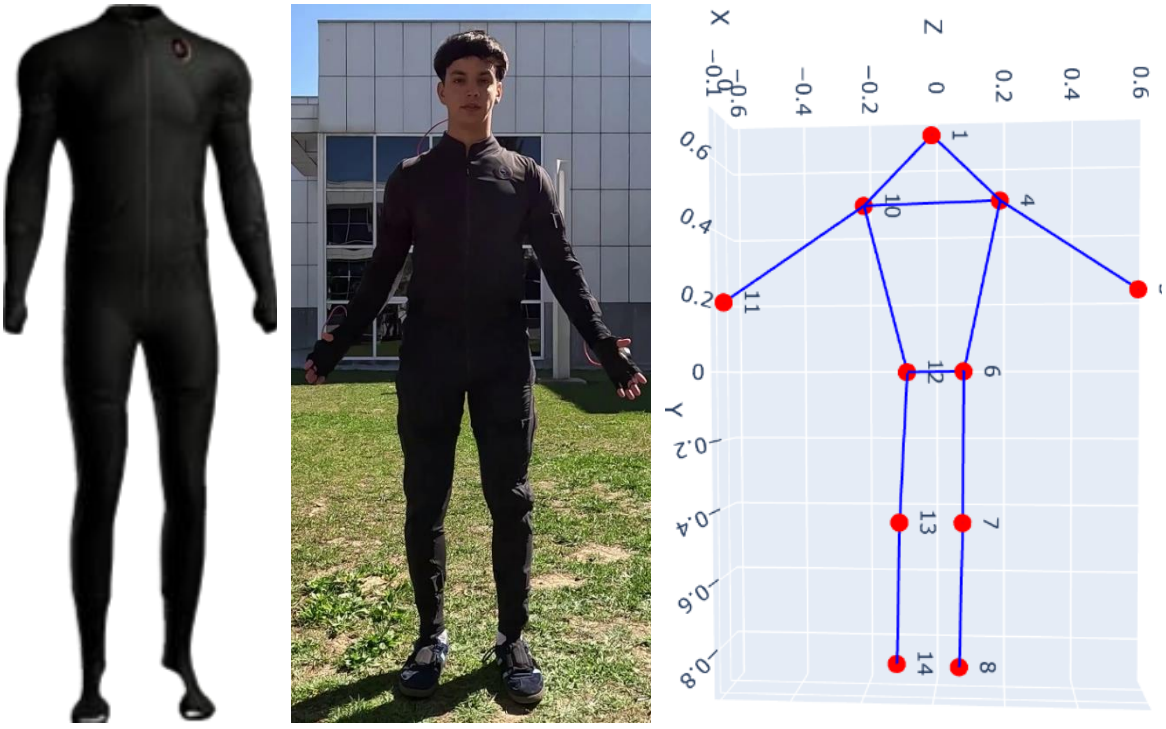

Figura 5. Vista general del sistema Rokoko Smartsuit Pro II: (izquierda) distribución de sensores inerciales sobre el traje; (centro) participante equipado con el traje durante una sesión de grabación; (derecha) representación en 3D de los keypoints corporales generados por el software a partir de los datos de movimiento

El software Rokoko Studio se instaló en un ordenador portátil que cumplía ampliamente con las especificaciones técnicas recomendadas por el fabricante. En concreto, el equipo contaba con sistema operativo Windows 11 Pro (versión 23H2), procesador Intel Core i5, y 8 GB de memoria RAM.

La versión utilizada fue Rokoko Studio 2.4.8.0, que incluye soporte nativo para dispositivos como el Smartsuit Pro II, los Smartgloves y el Coil Pro, además de incorporar mejoras de estabilidad y rendimiento en comparación con versiones anteriores (*Intuitive and Affordable Motion Capture Tools for Character Animation*, s. f.).

Durante la configuración inicial, todos los componentes del traje deben conectarse al ordenador mediante cables USB-C, lo cual permite tanto su detección y configuración en el software como la conexión a la red Wi-Fi del sistema. En esta etapa también se gestionan las actualizaciones de firmware necesarias para garantizar el funcionamiento correcto del hardware.

La red Wi-Fi utilizada fue una red privada 4G, generada mediante el punto de acceso personal de un teléfono móvil. En concreto, se empleó un Xiaomi Redmi 13C (Android, pantalla de 6.74"), lo que permitió configurar una red local (LAN) estable y de baja latencia. Esta red fue utilizada exclusivamente por el traje Smartsuit Pro II, los Smartgloves,

el ordenador portátil y el dispositivo Coil Pro. El uso de un teléfono móvil como punto de acceso presenta la ventaja de ser una solución portátil y autónoma.

Para mantener una conexión fluida durante las sesiones de grabación, se procuró que la intensidad de la señal Wi-Fi se mantuviera por encima del 95%. Para ello, el dispositivo móvil que generaba la red se colocaba siempre en el bolsillo del portador del traje, minimizando así la distancia entre el emisor y los dispositivos conectados. Además, el Smartsuit Pro II incorpora una función de reconexión automática, que permite restablecer rápidamente la conexión en caso de pérdida temporal de la señal.

5.1.2. Configuración del traje y toma de medidas corporales

Para garantizar un funcionamiento óptimo del sistema de captura de movimiento, fue necesario realizar una configuración personalizada para cada usuario. El software Rokoko Studio requiere introducir una serie de medidas antropométricas del perfil del actor (*Actor Profile*, s. f.), las cuales permiten ajustar el modelo esquelético interno y mejorar la precisión del cálculo de la pose. A continuación, se enumeran las medidas solicitadas, junto con su correspondiente descripción y referencia visual:

- Longitud del pie (Figura 6): Medida desde la parte posterior del talón (calcáneo) hasta la punta del segundo dedo del pie.

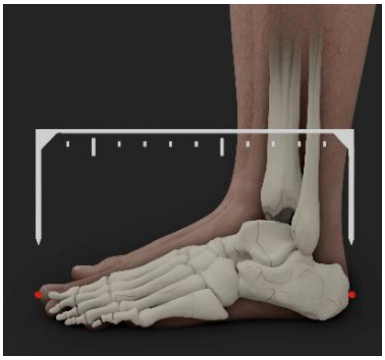

Figura 6. Longitud del pie

- Envergadura (Figura 7): Distancia entre las puntas de los dedos medios de ambas manos, con los brazos extendidos en posición de "T" (90 grados respecto al tronco).

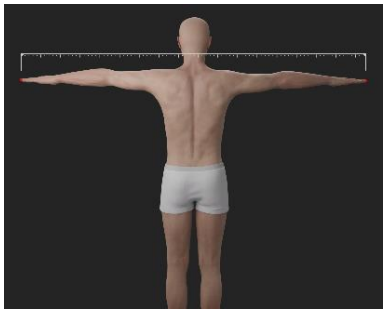

Figura 7. Envergadura

- Altura de los hombros (Figura 8): Medida desde el suelo hasta la escotadura esternal (el punto entre las clavículas, en la base del cuello).

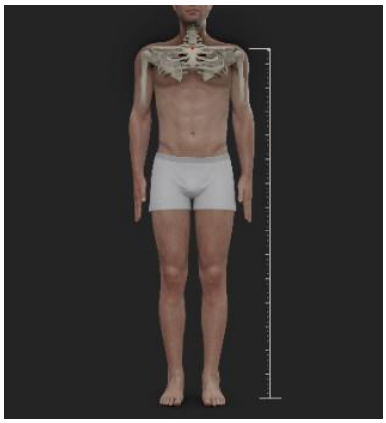

Figura 8. Altura de los hombros

- Anchura de hombros (Figura 9): Distancia entre los bordes externos izquierdo y derecho de la parte posterior de los hombros (punto de referencia: acromion).

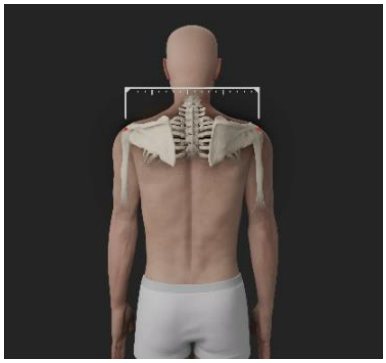

Figura 9. Anchura de los hombros

- Altura de la pelvis (Figura 10): Medida desde el suelo hasta la espina ilíaca anterosuperior (punto óseo prominente en la parte frontal de la pelvis).

Figura 10. Altura de la pelvis

- Anchura de la pelvis (Figura 11): Distancia entre las espinas ilíacas anterosuperiores izquierda y derecha, ubicadas sobre el centro de los muslos.

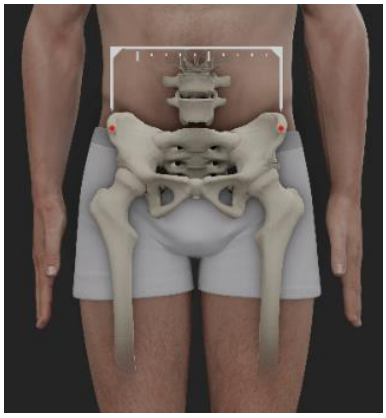

Figura 11. Anchura de la pelvis

- Altura de la rodilla (Figura 12): Medida desde el suelo hasta el cóndilo lateral de la rodilla (punto óseo en el lateral de la articulación).

Figura 12. Altura de la rodilla

- Longitud del brazo (Figura 13): Medida desde la punta del dedo medio hasta el olécranon (punto óseo en la parte posterior del codo).

Figura 13. Longitud del brazo

La precisión en la toma de medidas resulta crítica debido al funcionamiento del traje, basado en IMUs distribuidos estratégicamente a lo largo del cuerpo. Cada IMU registra aceleraciones lineales, velocidades angulares y, en algunos casos, campos magnéticos locales. A partir de estos datos, el sistema calcula la orientación de cada segmento corporal mediante integración numérica y estima la posición relativa de las articulaciones utilizando un modelo biomecánico personalizado. Para que esta reconstrucción sea fiable, el software necesita conocer con precisión las proporciones del usuario, ya que las posiciones se derivan de relaciones entre segmentos corporales enlazados.

Las medidas corporales introducidas definen parámetros fundamentales del modelo: las longitudes relativas entre sensores, las proporciones entre segmentos y las restricciones de movimiento articular. Errores en esta etapa pueden traducirse en estimaciones incorrectas, como extremidades desproporcionadas, orientaciones imprecisas o movimientos que no respetan las limitaciones biomecánicas humanas. Este tipo de desviaciones son especialmente relevantes en contextos donde se requiere alta fidelidad en la captura, como en entrenamiento de modelos de inteligencia artificial, análisis biomecánico o producción audiovisual, donde una acumulación de errores de pocos centímetros puede afectar de forma significativa la pose final.

Para minimizar estos errores, la toma de medidas se realizó siguiendo las recomendaciones oficiales de Rokoko. El usuario se mantuvo de pie, en una postura natural y relajada, usando ropa ajustada para evitar interferencias con los sensores. Las medidas se tomaron varias veces con cinta métrica flexible y se utilizó el valor promedio. Finalmente, los datos se introdujeron manualmente en Rokoko Studio en unidades de centímetros, verificando su coherencia antes de iniciar la calibración. Aunque el software permite pequeños ajustes posteriores, la calidad general del sistema depende en gran medida de la precisión en esta fase inicial.

Además de la configuración inicial del traje, antes de cada sesión se realiza una calibración que permite al sistema ajustar la posición y

orientación de los sensores en función de las medidas introducidas. Este proceso es esencial para que la reconstrucción del modelo corporal sea coherente con la anatomía real del usuario. Para comprobar la efectividad de la toma de medidas, se calcularon distancias entre keypoints en el modelo generado y se compararon con las medidas reales registradas. Esta verificación objetiva aseguró que las proporciones y relaciones espaciales fueran correctas, evitando depender únicamente de la inspección visual.

En conclusión, una toma de medidas rigurosa y cuidadosa fue un requisito indispensable para asegurar el éxito en las fases posteriores de calibración y captura de datos, minimizando posibles fuentes de error y asegurando una representación fiable y realista del movimiento humano.

5.1.3. Uso del geoposicionamiento Coil Pro

Con el objetivo de mejorar la precisión espacial del sistema de captura de movimiento, se integró el dispositivo Coil Pro de Rokoko. Este dispositivo de geoposicionamiento magnético proporciona información de la orientación y posición global del usuario en el espacio, complementando los datos relativos proporcionados por las IMUs del traje Rokoko Smartsuit Pro II, y mejorando así la estabilidad en sesiones de captura prolongadas. Se puede ver visualizado en el software de Rokoko Studio en la Figura 14.

Figura 14. Aspecto físico del Coil Pro (izquierda), su integración en el sistema de caputra (centro) y su visualización en el software de Rokoko Studio.

El Coil Pro es un generador campo magnético de baja frecuencia y alta estabilidad en su entorno inmediato, presentado en la Figura 14. Este campo es detectado por los sensores magnetométricos integrados en los Smartgloves, que permiten calcular de manera precisa su posición tridimensional (coordenadas X, Y, Z) y su orientación (pitch, roll, yaw) relativa al Coil Pro (*Coil Pro FAQs*, s. f.). El origen del sistema de

coordenadas se establece en la proyección del centro del Coil Pro sobre el suelo, siendo este el punto de referencia espacial desde el cual se calcula la posición de los keypoints.

A diferencia de otras tecnologías de posicionamiento, como el GPS, cuyas ondas electromagnéticas de gran longitud no son adecuadas para entornos cerrados, o los sistemas de visión óptica, que requieren infraestructuras costosas y voluminosas, el Coil Pro ofrece una solución compacta, precisa y adaptable a distintos entornos.

Para las sesiones de grabación, el Coil Pro se conectó a la red eléctrica con una alargadera, ya que requiere una tensión elevada incompatible con baterías portátiles. Antes de comenzar cada sesión de captura, se efectuó una calibración inicial denominada Zero Calibration desde la interfaz de Rokoko Studio. Este proceso permite establecer la referencia de orientación y corregir posibles sesgos en las mediciones magnéticas, siendo fundamental para garantizar la calidad y estabilidad de los datos obtenidos.

Durante el proceso de utilización del Coil Pro se identificaron varios desafíos técnicos, los cuales se han abordado siguiendo las recomendaciones de la documentación oficial (*Coil Pro - Known Issues*, s. f.):

- Interferencias electromagnéticas generadas por la proximidad de objetos metálicos grandes, cables de alta tensión o estructuras de acero, que afectan negativamente la precisión espacial detectada por los guantes. Para mitigar este efecto, se recomienda situar el Coil Pro a una distancia mínima de 1,5 metros de dichos objetos, prefiriendo entornos exteriores libres de metales.
- Desincronización entre el Coil Pro y los guantes, manifestada en comportamientos erráticos como posiciones flotantes o desplazamientos no realistas de las manos. Esta situación se solventa mediante el reinicio periódico del dispositivo y la recalibración del usuario.
- Dificultades en la detección inmediata del Coil Pro al iniciar Rokoko Studio. Para solucionarlo, se aconseja desconectar brevemente la alimentación del dispositivo y volver a conectarla antes de reintentar la detección en el software.
- Desconexiones intermitentes en ambientes con alta interferencia electromagnética, problema conocido en ciertas unidades. Para minimizarlo, se recomienda mantener actualizado el firmware tanto del Coil Pro como del Smartsuit Pro II.
- Para asegurar la calidad en la captura de datos, se estableció posicionar el Coil Pro a una altura aproximada de 72 centímetros sobre el suelo, realizar recalibraciones periódicas en sesiones prolongadas, y llevar a cabo grabaciones de prueba antes de cada sesión oficial para verificar la consistencia en la posición y orientación detectadas por el sistema.

La integración adecuada del Coil Pro ayudó a evitar el drift durante las capturas, garantizando así una mayor fiabilidad y precisión en los datos desarrollados en el proyecto.

5.1.4. Entorno y condiciones de grabación

La elección del entorno de grabación es un aspecto clave en la definición del dataset experimental, ya que las condiciones físicas del espacio donde se lleva a cabo la captura de movimiento influyen directamente en la precisión del sistema, la calidad de los datos obtenidos y la validez de su uso posterior en tareas de entrenamiento y análisis. En este proyecto, se ha optado por realizar las sesiones de captura al aire libre, en lugar de en espacios cerrados como laboratorios o salas técnicas. Esta decisión se basa en la necesidad de minimizar las interferencias electromagnéticas provocadas por objetos metálicos presentes en entornos interiores, tales como mesas, ordenadores, cableado eléctrico o estructuras metálicas. Además, seleccionar entornos diversos contribuye a introducir mayor variabilidad en la literatura existente, que suele centrarse en contextos altamente controlados.

Se han tomado las siguientes consideraciones para asegurar la calidad y estabilidad del entorno de grabación:

- Posición del Coil Pro: Se situó sobre una mesa no metálica de aproximadamente 72 cm de altura, actuando como punto de referencia central desde el cual se proyectan los keypoints y se define el sistema de coordenadas global, evitando interferencias magnéticas.
- Variabilidad temporal: Las sesiones de captura se distribuyeron en varios días distintos, con el fin de registrar variabilidad en las condiciones de iluminación y meteorología (luz solar directa, nubes, temperatura, orientación solar).
- Diversidad espacial: Se planificaron grabaciones en diferentes ubicaciones y con distintas orientaciones de cámara, de forma que se alteraran el fondo y la incidencia de la luz, enriqueciendo la variabilidad visual del conjunto de datos.
- Condición del terreno: Se seleccionó un suelo plano, en este caso césped natural, para evitar inclinaciones o irregularidades que pudieran comprometer la precisión del Smartsuit Pro II en los cálculos de orientación y posición.
- Entorno libre de interrupciones: Se eligieron zonas poco transitadas, minimizando la presencia de personas ajenas en el campo de visión de las cámaras y evitando así ruido visual e interferencias que pudieran afectar a la anotación o al entrenamiento de modelos de visión artificial.

En conjunto, estas decisiones permitieron establecer un entorno de grabación controlado pero representativo, que favorece la calidad del dataset sin sacrificar la variabilidad natural del mundo real.

5.1.5. Selección de participantes

Los participantes fueron seleccionados con el objetivo de aportar variabilidad visual y de movimientos al conjunto de datos, asegurando que las grabaciones representen distintos perfiles físicos. Esta diversidad es clave para mejorar la generalización de los modelos de aprendizaje automático y evitar el sobreajuste a condiciones específicas.

En total, el conjunto de datos incluyó grabaciones correspondientes a ocho sujetos distintos, entendiendo cada sujeto como una combinación única de persona, morfología, género y apariencia visual (vestimenta). De estos ocho casos, cinco correspondieron a un mismo individuo (el autor del proyecto), grabado en diferentes días y con distintos atuendos, lo que introdujo una variabilidad visual significativa sin comprometer la homogeneidad en la ejecución de los movimientos. Esta estrategia permitió combinar la consistencia gestual con la variabilidad superficial, aspecto especialmente útil en tareas relacionadas con la detección o clasificación basada en características visuales.

Las principales características de los ocho sujetos considerados se pueden reflejar en la Tabla 1:

Tabla 1. Tabla con los sujetos participantes en el dataset y sus principales características

Sujeto	Persona	Género	Altura (cm)	Color
1	1	Masculino	175	Azul
2	1	Masculino	175	Gris
3	2	Masculino	178	Rojo
4	3	Femenino	158	Blanco
5	1	Masculino	175	Añil
6	4	Femenino	172	Verde
7	5	Masculino	175	Beige
8	6	Masculino	175	Rojo

La selección incluyó sujetos con diferentes complexiones físicas, colores de ropa, tipos de prendas y géneros. Concretamente, dos de los ocho sujetos fueron mujeres, lo que permitió incorporar cierta diversidad de género en el conjunto de datos. Asimismo, la elección de ropa con tonalidades y contrastes variados contribuyó a evaluar cómo respondía el sistema de captura y proyección ante distintos patrones visuales.

5.1.6. Tipos de movimientos a capturar

Uno de los objetivos fundamentales en la definición del conjunto de datos experimental fue capturar un repertorio de movimientos representativo de la actividad cotidiana humana, abarcando la mayoría de los patrones de movimiento presentes en contextos reales. Para ello, se diseñó un conjunto estructurado de escenas que los participantes debían ejecutar durante las sesiones de grabación, de forma que el conjunto de datos resultante combinara variabilidad gestual, riqueza postural y naturalidad en las acciones.

Cada uno de los ocho sujetos definidos previamente realizó un total de 14 escenas diferentes, seleccionadas con el objetivo de cubrir un amplio espectro de movimientos articulares, interacciones físicas y expresiones corporales habituales. Las acciones se definieron cuidadosamente para maximizar la diversidad sin comprometer la estabilidad técnica del sistema de captura basado en sensores inerciales.

A continuación, se describen las escenas que conformaron el conjunto de movimientos capturados:

1. Caminar: el actor camina en línea recta, en círculos y en zigzag, a diferentes velocidades y en distintos sentidos (de frente, de espaldas, lateralmente).
2. Agacharse: se realizan movimientos de flexión de piernas y tronco, en los que el actor se agacha y se incorpora de forma repetida, utilizando diferentes apoyos y gestos.
3. Estiramientos: el actor realiza diversos ejercicios de estiramiento, incluyendo movimientos de brazos, piernas, cuello y espalda, en diferentes direcciones y posturas.
4. Hablar: se simula una conversación en la que el actor mira a distintos puntos y gesticula de manera variable con manos y rostro, como en una interacción social real.
5. Vestirse: el actor simula acciones de ponerse y quitarse prendas, como una chaqueta, sudadera o gafas, incluyendo gestos como subirse una cremallera o ajustarse las mangas.
6. Direcciones: el actor representa distintas situaciones en las que indica direcciones o explicaciones físicas mediante gestos amplios con los brazos y el cuerpo.
7. Teléfono: se simula la acción de coger una llamada, mantener una conversación breve y colgar, incluyendo gestos asociados como caminar, asentir o cambiar de mano.
8. Fumar: el actor simula encender un cigarro, llevarlo a la boca, inhalar y exhalar, con diferentes posturas de brazo y manos, y finalmente apagarlo.

9. Foto: se representa el acto de sacar un móvil, realizar fotografías frontales y hacia el entorno, incluyendo gestos como enfocar, encuadrar y revisar imágenes.
10. Posar: el actor realiza distintas poses corporales como si fuese un modelo ante la cámara, variando la orientación del cuerpo, la distribución del peso y la expresión.
11. Música: se simula la acción de escuchar música con auriculares, incluyendo gestos de disfrute, movimiento rítmico del cuerpo y expresión facial relajada o animada.
12. Bailar: el actor representa diferentes tipos de baile, alternando movimientos suaves y enérgicos, desplazamientos cortos, balanceo y uso de brazos y caderas.
13. Sentarse: se realizan varias acciones de sentarse y levantarse, con diferentes niveles de velocidad y estilo, además de pequeños gestos realizados mientras el actor permanece sentado.
14. Objeto: el actor interactúa con un objeto pequeño (como una pelota o un bolígrafo), mostrándolo a cámara, lanzándolo, recogiéndolo o manipulándolo con una o ambas manos.

Cabe destacar que, debido a limitaciones inherentes al sistema de captura empleado (basado en sensores inerciales, IMUs), se excluyeron aquellas acciones que implicaran el despegue simultáneo de ambos pies del suelo, como correr o saltar. Este tipo de movimientos podía provocar errores de calibración o pérdidas temporales de referencia, comprometiendo la precisión de los datos capturados y su utilidad en fases posteriores de análisis.

En conjunto, el conjunto de escenas definidas proporcionó un equilibrio entre realismo, cobertura gestual y viabilidad técnica, garantizando un conjunto de datos rico y variado, pero compatible con las restricciones del sistema de captura empleado.

5.1.7. Especificaciones de las cámaras

Para la captura del conjunto de escenas definidas en este proyecto, se ha empleado un sistema de grabación multicámara formado por tres unidades GoPro HERO11 Black (*GoPro HERO11 Black (cámara deportiva y subacuática)*, s. f.). Estas cámaras han sido seleccionadas por su alta calidad de imagen, su robustez para entornos exteriores y su facilidad de uso en configuraciones sincronizadas.

Cada una de las tres cámaras se configuró para grabar a una resolución de 4K con perfil de lente lineal (sin distorsión de ojo de pez) y a una frecuencia de 60 fotogramas por segundo (fps). Posteriormente, durante el procesamiento de los datos, se extrajeron los frames a 30 fps,

con el objetivo de igualar la frecuencia de muestreo de los datos inerciales obtenidos del traje Rokoko Smartsuit Pro II y facilitar la sincronización entre ambas fuentes de información.

En cuanto a la disposición espacial, las cámaras se colocaron estratégicamente alrededor del sujeto, manteniendo una distancia uniforme de aproximadamente 2 a 3 metros en cada caso. Las posiciones concretas se pueden visualizar en la Figura 15 y se desarrollan a continuación:

Figura 15. Disposición espacial de las cámaras en el entorno de grabación

- Cámara principal (gopro3_principal): situada justo delante del sujeto, montada sobre el dispositivo Coil Pro, define el eje de referencia central del sistema.
- Cámara lateral (gopro2_lateral): colocada a la izquierda del sujeto, en posición perpendicular a su dirección frontal, permite capturar el perfil completo durante la ejecución de los movimientos.
- Cámara trasera (gopro4_trasera): ubicada detrás y ligeramente hacia la derecha, proporciona una vista posterior oblicua que resulta útil para el análisis tridimensional y la evaluación de la simetría de los gestos.

Para asegurar una sincronización perfecta entre las cámaras, se utilizó un dispositivo de activación remota, GoPro Remote, consistente en un mando que permite iniciar y detener la grabación en las tres cámaras simultáneamente mediante un único botón. Este sistema garantiza que todas las cámaras comiencen y terminen la grabación en el mismo instante, lo que permite la sincronización entre secuencias y facilita su posterior procesamiento conjunto.

5.2. GENERACIÓN DEL DATASET

La generación del conjunto de datos constituye una fase central del proyecto, en la que se combinan tanto la captura efectiva de los datos como su preparación y estructuración posterior. Esta etapa se alinea con las fases de obtención y comprensión de los datos dentro de la metodología CRISP-DM, y tiene como objetivo producir un dataset coherente, sincronizado y técnicamente válido para su posterior uso en tareas de evaluación y análisis.

En primer lugar, se evaluaron distintas estrategias de calibración que permitieran obtener los distintos parámetros de las cámaras. A partir de estas pruebas se realizaron la calibración de las cámaras mediante patrones de reconocimiento, así como la estimación de la posición relativa entre ellas, lo que permitió capturar escenas desde múltiples puntos de vista de forma consistente.

Posteriormente, se llevó a cabo la grabación multicámara y la correspondiente sincronización con los datos inerciales generados por el traje y los guantes. A continuación, se procedió a la extracción de keypoints y secuencias de vídeo, y a la alineación temporal entre ambos. Finalmente, se trató de proyectar los keypoints sobre las imágenes 2D y se estructuró el conjunto de datos con el formato necesario para su uso en modelos de aprendizaje automático.

Esta fase no solo permitió capturar y sincronizar la información visual e inercial, sino que dejó todos los datos preparados para la posterior estructuración y análisis del dataset.

5.2.1. Estudio de distintas estrategias de calibración

En este apartado, se presentan las distintas estrategias estudiadas y exploradas relacionadas con la calibración de las cámaras, con el objetivo de calcular los parámetros intrínsecos y extrínsecos de las cámaras GoPro utilizadas durante la grabación.

Por un lado, con un patrón de reconocimiento del tipo checkerboard, se puede obtener la matriz de calibración interna, que incluye parámetros como la distancia focal, el punto principal y los coeficientes de distorsión óptica. Estos valores son necesarios para corregir las deformaciones introducidas por la lente de la GoPro. La calibración intrínseca resulta fundamental para garantizar que las proyecciones geométricas se correspondan fielmente con la realidad física observada en los vídeos.

Por otro lado, se consideró una calibración extrínseca, especialmente relevante en el contexto de capturas multivista. Esta técnica permite determinar la posición relativa entre cámaras utilizando el mismo patrón checkerboard, a partir del cual se calculan la matriz de rotación y la matriz

de traslación que definen la transformación espacial entre los distintos dispositivos ópticos.

Estas estrategias sentaron las bases necesarias para las pruebas de calibración posteriores y la correcta alineación de los datos visuales en el proceso de generación del conjunto de datos.

5.2.1.1. Prueba de calibración de la cámara con patrón de reconocimiento

En esta tarea, se llevó a cabo una prueba para evaluar distintos parámetros de la cámara en relación con la visión de un checkerboard 9x6. El procedimiento consistió en grabar videos desde diversas perspectivas con la GoPro sobre un trípode, como se muestra en la Figura 16, y extraer frames de los mismos, ya que el principal interés residía en obtener la visión de la cámara al grabar videos, no en realizar fotografías.

Figura 16. Ejemplo de grabación del chessboard con la GoPro

Durante este proceso, se registraron las distancias en los ejes de coordenadas (x, y, z) respecto al punto de referencia del patrón de reconocimiento, así como la distancia de la cámara al punto de referencia. En total, se capturaron 6 frames de video para llevar a cabo el análisis. En la Figura 17 se puede observar la representación gráfica de los ejes y medidas que tomaron en cuenta.

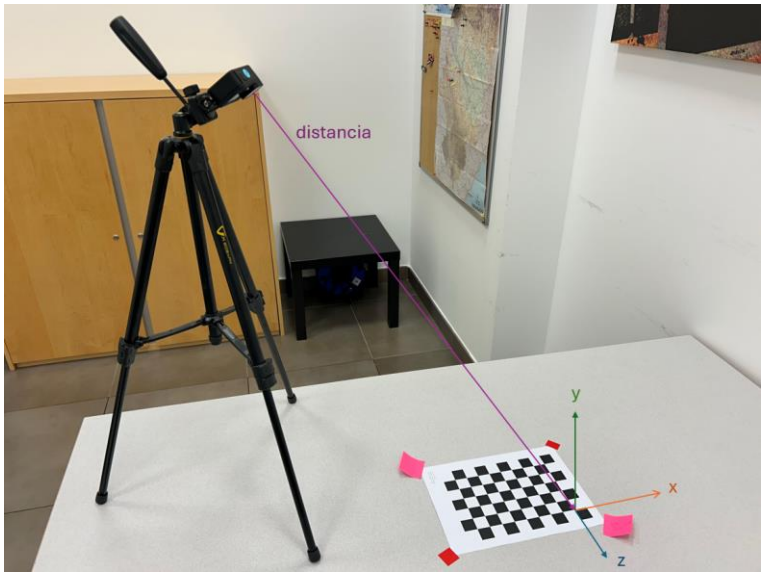

Figura 17. Representación gráfica de las medidas que se tomaron

Una vez obtenidos los frames, se procedió con el procesamiento de las imágenes. En primer lugar, se calculó la matriz intrínseca de la cámara, la cual describe las características internas de la cámara que afectan la proyección de los puntos 3D del mundo real en las imágenes. Para obtenerla, como 6 frames no son suficientes, se grabó a parte un video enfocando al patrón de reconocimiento desde distintos ángulos y del cuál se obtuvieron 16 frames, con los que se obtuvo la matriz intrínseca:

$$M = \begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 2005.53 & 0 & 1916.13 \\ 0 & 2003.03 & 1112.09 \\ 0 & 0 & 1 \end{bmatrix}$$

Cada número tiene el siguiente significado:

- Los valores de la diagonal son las longitudes focales de la cámara en los ejes x e y, respectivamente, expresadas en magnitudes de píxeles. Estas longitudes focales determinan la magnificación de la imagen en función de la distancia entre la cámara y el objeto.
- Los valores de la tercera columna son las coordenadas del centro de la imagen en los ejes x e y, que corresponden a los puntos donde el eje óptico de la cámara corta la imagen.

Con las mismas imágenes con las que se sacó la matriz intrínseca, se calcularon los coeficientes de distorsión que corrigen las aberraciones ópticas causadas por la lente de la cámara. Estos coeficientes son importantes para obtener una proyección precisa del mundo real en la imagen:

$$D = [k_1, k_2, p_1, p_2, k_3] = [8.77x10^{-3}, 3{,}78x10^{-4}, 3.21x10^{-3}, -1.39x10^{-3}, 1.16x10^{-3}]$$

Después de obtener la matriz intrínseca y los coeficientes de distorsión, se procedió a calcular los parámetros de rotación y translación para cada una de las imágenes. Estos parámetros describen cómo el

sistema de coordenadas del mundo real (en este caso, el patrón de reconocimiento) se transforma en el sistema de coordenadas de la cámara. Para ello, se calculó previamente el tamaño de los cuadrados del patrón de reconocimiento: 2.42 cm.

Con estos parámetros, se intentó inicialmente proyectar un punto rojo con las medidas x, y, z calculadas en relación con las casillas del patrón de reconocimiento. Sin embargo, este intento no funcionó correctamente porque las coordenadas del punto debían estar definidas en el sistema de coordenadas de la cámara, no en las coordenadas del patrón de reconocimiento.

Para verificar la precisión de las mediciones, se creó un programa para calcular la distancia desde el punto de referencia del patrón de reconocimiento a la cámara. Este cálculo se comparó con la distancia previamente medida. Aunque el resultado obtenido fue muy similar al medido manualmente, en los seis casos el programa sacó una distancia ligeramente mayor: un error absoluto medio de 0.463 cm y un error relativo medio de 0.56%.

Para comprobar que la medición estuviera siendo realizada respecto al punto de referencia correcto, se proyectó el punto de referencia en la imagen obtenida (Figura 18), y se comprobó que efectivamente coincidía con el punto al que se había hecho la medición.

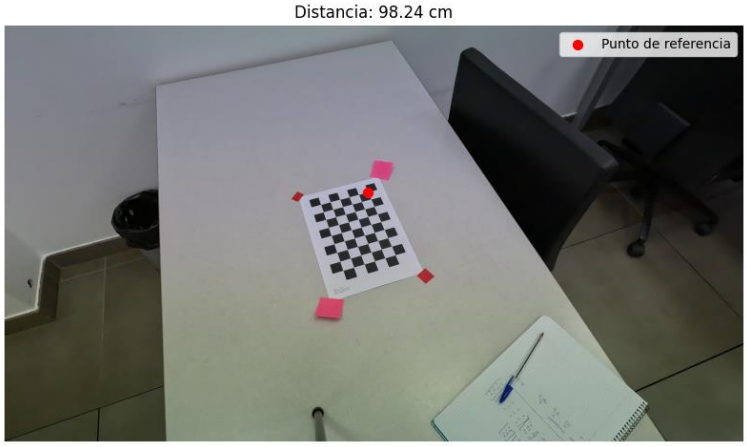

Figura 18. Proyección del punto de referencia y la distancia del foco al punto

Posteriormente, se diseñó un programa que medía las distancias en coordenadas x, y y z desde la cámara hasta el punto de referencia, utilizando como referencia el sistema de coordenadas del patrón de reconocimiento. Los resultados obtenidos fueron similares a las mediciones manuales y, por ejemplo, en el caso de la altura (eje y) se obtuvo un error muy similar al de la distancia calculado anteriormente: un error absoluto medio de 0.295 cm y un error relativo medio de 0.40%.

Por lo tanto, los resultados obtenidos demostraron ser correctos y coherentes con las mediciones manuales, lo que valida la metodología empleada en la prueba. La ligera discrepancia en las mediciones podría

deberse a pequeños errores de medición o al hecho de que el foco de la cámara esté ligeramente desplazado hacia el interior del dispositivo, lo que afecta la referencia espacial utilizada en los cálculos.

5.2.1.2. Prueba de estimación de la posición relativa entre cámaras

En esta prueba, se añadió una segunda cámara al sistema de calibración (véase la Figura 19). Antes de su uso, se realizó su calibración siguiendo el mismo procedimiento que con la primera cámara (apartado 5.2.1.1), obteniendo así la matriz intrínseca y los coeficientes de distorsión. Dado que ambas cámaras son del mismo modelo y cuentan con los mismos ajustes de video, los valores obtenidos fueron similares.

Figura 19. Ejemplo de grabación del patrón de reconocimiento con dos cámaras

Tras la calibración, se colocaron ambas cámaras sobre un trípode, apuntando al patrón de reconocimiento y formando un ángulo aproximado de 90°. Se midió la distancia entre ellas, obteniendo un valor de aproximadamente 60 cm. En la Figura 20 se muestran gráficamente las medidas tomadas. Al igual que en la prueba anterior, en lugar de capturar fotografías, se grabó un video estático del cual se extrajo un frame por cada cámara.

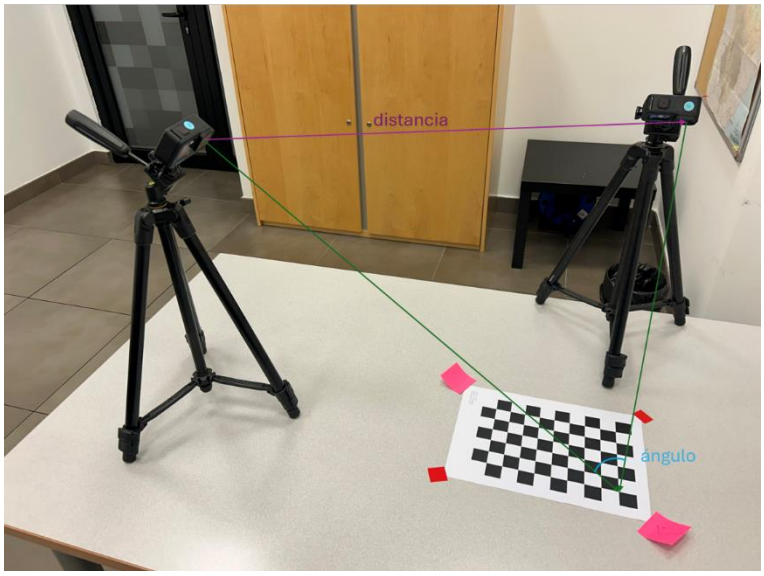

Figura 20. Representación gráfica de las medidas que se van a tomar

A partir de los frames obtenidos, se desarrolló un programa para calcular la matriz de rotación R y el vector de traslación T de la segunda cámara con respecto a la primera.

- Matriz de rotación R: Describe cómo está orientada la segunda cámara con respecto a la primera. Esta matriz transforma las coordenadas del sistema de la primera cámara al sistema de la segunda, indicando la rotación entre ambos marcos de referencia.

$$R = \begin{bmatrix} r_{11} & r_{12} & r_{13} \\ r_{21} & r_{22} & r_{23} \\ r_{31} & r_{32} & r_{33} \end{bmatrix} = \begin{bmatrix} 0.05 & 0.77 & -0.63 \\ -0.73 & 0.46 & 0.50 \\ 0.68 & 0.44 & 0.59 \end{bmatrix}$$

- Vector de traslación T: Representa la posición relativa de la segunda cámara con respecto a la primera en el espacio tridimensional.

$$T = [t_x, t_y, t_z] = [52.76, -21.6, 19.56]$$

A partir de estos parámetros, se calcularon la distancia y el ángulo de rotación entre cámaras:

- Distancia entre cámaras: 60.27 cm
- Ángulo de rotación entre cámaras: 86.94°

Los valores obtenidos concordaban con las mediciones manuales (60 cm y 90° estimados visualmente), lo que valida el procedimiento realizado.

Este método puede aplicarse en la grabación multicámara con el traje Rokoko. Si se coloca un patrón de reconocimiento en el suelo, es posible determinar la posición y orientación exacta de cada cámara respecto a la cámara principal, permitiendo mejorar la precisión en la reconstrucción de la pose.

5.2.2. Grabación multicámara y sincronización de secuencias

La fase de grabación se llevó a cabo a lo largo de varios días, utilizando tanto el traje de captura de movimiento como el sistema multicámara compuesto por tres cámaras. Durante todas las sesiones se siguieron cuidadosamente las especificaciones previamente descritas en cuanto al entorno de grabación, los sujetos implicados, las acciones a ejecutar y la disposición de las cámaras, con el objetivo de garantizar la coherencia y reproducibilidad de los datos obtenidos.

Previo al inicio de cada grabación, se colocaba un patrón de reconocimiento (checkerboard) en el centro del espacio de grabación, de manera que fuera visible por las tres cámaras simultáneamente. Esta configuración permitía calcular posteriormente la posición relativa de las cámaras secundarias respecto a la cámara principal, asegurando una correcta alineación espacial entre las distintas vistas.

Además, antes de cada escena se procedía al reinicio del sistema Coil Pro y a la calibración del sujeto que portaba el traje de captura. Este paso era esencial para garantizar la precisión de los datos de movimiento recogidos y su sincronización con las imágenes obtenidas por las cámaras.

5.2.3. Extracción entre keypoints e imágenes

Una vez finalizadas las grabaciones, se disponía de dos tipos de datos: los vídeos capturados por las tres cámaras y los datos de movimiento recogidos por el sistema Rokoko a través del traje de captura.

Los datos procedentes del sistema Rokoko se exportaron inicialmente en formato .bvh (Biovision Hierarchy File), ampliamente utilizado en animación y captura de movimiento. Sin embargo, este formato no es el más adecuado para su procesamiento en entornos como Python, por lo que fue necesario convertirlo a un formato más manejable, concretamente .csv. Para ello, se utilizó un repositorio de GitHub desarrollado por el usuario tekulvw (Tekulve, 2016/2025).

El archivo .csv resultante contenía una fila por cada fotograma de la grabación, y columnas que representaban las coordenadas tridimensionales (x, y, z) de cada uno de los keypoints del cuerpo humano.

Paralelamente, se procesaron los vídeos grabados por las cámaras GoPro. Como se ha comentado anteriormente, aunque estaban registrados a 60 fotogramas por segundo, se desarrolló un script en Python utilizando la librería OpenCV (cv2) para reducir la frecuencia a 30

frames por segundo. Este ajuste permitió igualar la frecuencia de muestreo de los keypoints generados por Rokoko, evitando así problemas de desalineación temporal y facilitando la sincronización entre ambos tipos de datos.

5.2.4. Sincronización entre keypoints y frames

Previo a la proyección de los keypoints tridimensionales sobre los frames extraídos del vídeo, es necesaria la sincronización temporal entre ambos tipos de datos. Para ello, se diseñó una estrategia de sincronización basada en la identificación de un gesto visual claramente reconocible al inicio de la grabación: una palmada ejecutada por el usuario. Esta acción genera es fácilmente identificable tanto en el vídeo (por el contacto repentino de las manos) como en los datos inerciales registrados por el Smartsuit Pro II.

Desde el punto de vista analítico, se trata de determinar el instante exacto en el que las manos del usuario están más próximas entre sí. Para ello, se utilizaron las coordenadas tridimensionales (x, y, z) de los sensores inerciales situados en ambas manos y se calculó la distancia euclídea entre ellas para cada fila del archivo .csv generado a partir del fichero .bvh.

La distancia euclídea entre dos puntos P1=(x1,y1,z1) y P2=(x2,y2,z2) se define como:

$$d(P_1, P_2) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2 + (z_2 - z_1)}$$

Este cálculo permitió identificar la fila del fichero .csv en la que dicha distancia alcanzaba su valor mínimo, lo que coincide con el momento de contacto de la palmada. Por otro lado, se inspeccionaron los frames del vídeo para localizar manualmente aquel en el que visualmente se observa el gesto de la palmada. En la Figura 21 se puede visualizar como se detectó la palmada tanto en los frames como en las filas del csv. En la imagen se visualiza el momento exacto de la palmada, y en la gráfica de las distancias euclídeas entre las coordenadas de ambas manos se observa un valor mínimo. La fila en la que se alcanza ese mínimo corresponde con el dato registrado por Rokoko que coincide con el frame en el que se produce la palmada, consiguiendo la sincronización de ambos tipos de datos.

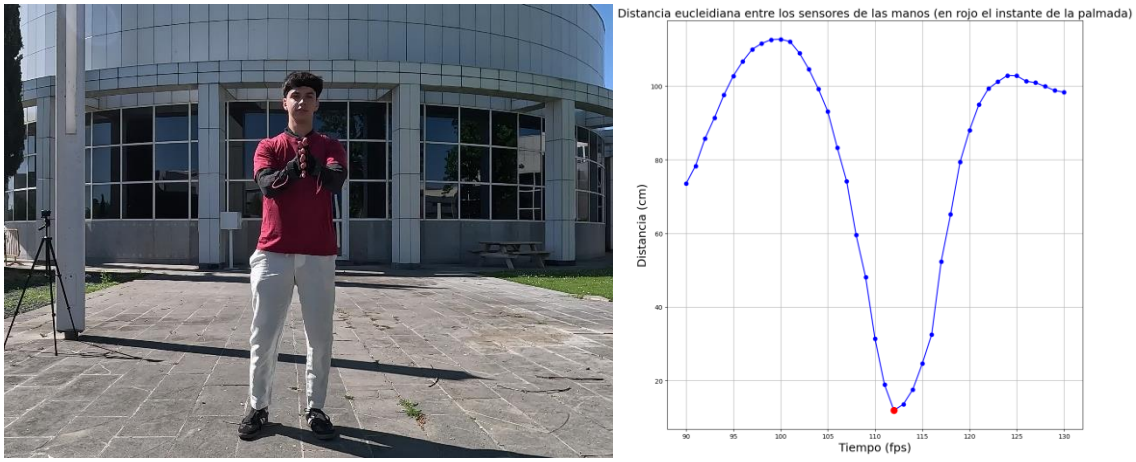

Figura 21. Visualización de la palmada en el frame (izquierda) y representación gráfica del instante de la palmada a partir de la distancia eucleideana entre los sensores de las manos (derecha)

Una vez identificados ambos momentos de referencia —la fila mínima del .csv y el frame correspondiente—, se estableció la correspondencia temporal entre el resto de las filas y los frames. Dado que ambos conjuntos de datos estaban ya alineados en frecuencia 30 fps (véase apartado 5.2.3), a partir del punto de sincronización inicial fue posible asociar directamente cada imagen a su correspondiente fila de keypoints, manteniendo una relación uno a uno.

Este paso es necesario dentro del pipeline de procesamiento de datos, ya que garantiza que las futuras proyecciones gráficas de los keypoints sobre las imágenes se realicen de forma precisa, evitando desajustes que podrían comprometer la calidad de la visualización o el entrenamiento de modelos basados en visión por computador.

5.2.5. Proyección de keypoints sobre los frames

Una vez sincronizados los keypoints tridimensionales y los frames extraídos del vídeo, se procedió a realizar la proyección de los puntos 3D sobre las imágenes. El objetivo de este proceso es representar visualmente, sobre cada fotograma, la posición estimada de las principales articulaciones del cuerpo humano, obtenidas a partir de los sensores inerciales del Rokoko Smartsuit Pro II.

Para llevar a cabo esta tarea, se utilizaron los parámetros de calibración de las cámaras, previamente calculados, y los cuales se han explicado con más detalle anteriormente (véase el apartado 5.2.1).

Los keypoints, inicialmente en el sistema de coordenadas del traje, fueron transformados aplicando el vector de traslación, obtenido a partir de la posición de las cámaras respecto al origen de coordenadas. Posteriormente, se invirtieron los ejes de las coordenadas según el siguiente cambio de base, necesario para adaptar el sistema del traje al

sistema de cámara (donde el eje Z apunta hacia adelante, el Y hacia abajo y el X hacia la derecha):

$$Coordenada\ transformada = \begin{bmatrix} Z \\ -Y \\ X \end{bmatrix} + \vec{t}_{cam}$$

Con las coordenadas transformadas, se empleó la función *cv2.projectPoints()* de OpenCV (*OpenCV: Camera Calibration and 3D Reconstruction*, s. f.), la cual implementa el modelo de cámara pinhole, para obtener las coordenadas del punto proyectado en el plano de imagen. Este modelo asume una cámara ideal sin lentes ni distorsión, donde los rayos de luz pasan a través de un único punto (el pinhole) y se proyectan en un plano de imagen, generando una proyección en perspectiva. Matemáticamente, la relación entre un punto 3D en el sistema de la cámara y su proyección 2D en píxeles puede expresarse como:

$$s \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = K \cdot \left[R \middle| \vec{t}\right] \cdot \begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix}$$

- u, y v corresponden a los píxeles de la imagen 2D.
- X, Y y Z corresponden al punto en coordenadas 3D que se quiere representar.
- K es la matriz intrínseca (véase en el apartado 5.2.1.1).
- $[R|\vec{t}]$ es la matriz de transformación, que incluye rotación y traslación (véase en el apartado 5.2.1.2). En esta primera prueba se consideró identidad para R y nulo para $\vec{t}$, asumiendo que los puntos ya estaban expresados en el sistema de cámara.
- s es un factor de escala homogéneo que permite igualar coordenadas homogéneas. Su valor depende de la profundidad Z del punto en la cámara y los parámetros intrínsecos/extrínsecos.

Finalmente, se representaron los puntos proyectados sobre el fotograma correspondiente utilizando la biblioteca OpenCV, lo que permitió visualizar directamente la localización estimada de las articulaciones. En la Figura 22 se puede observar de manera gráfica el proceso anterior:

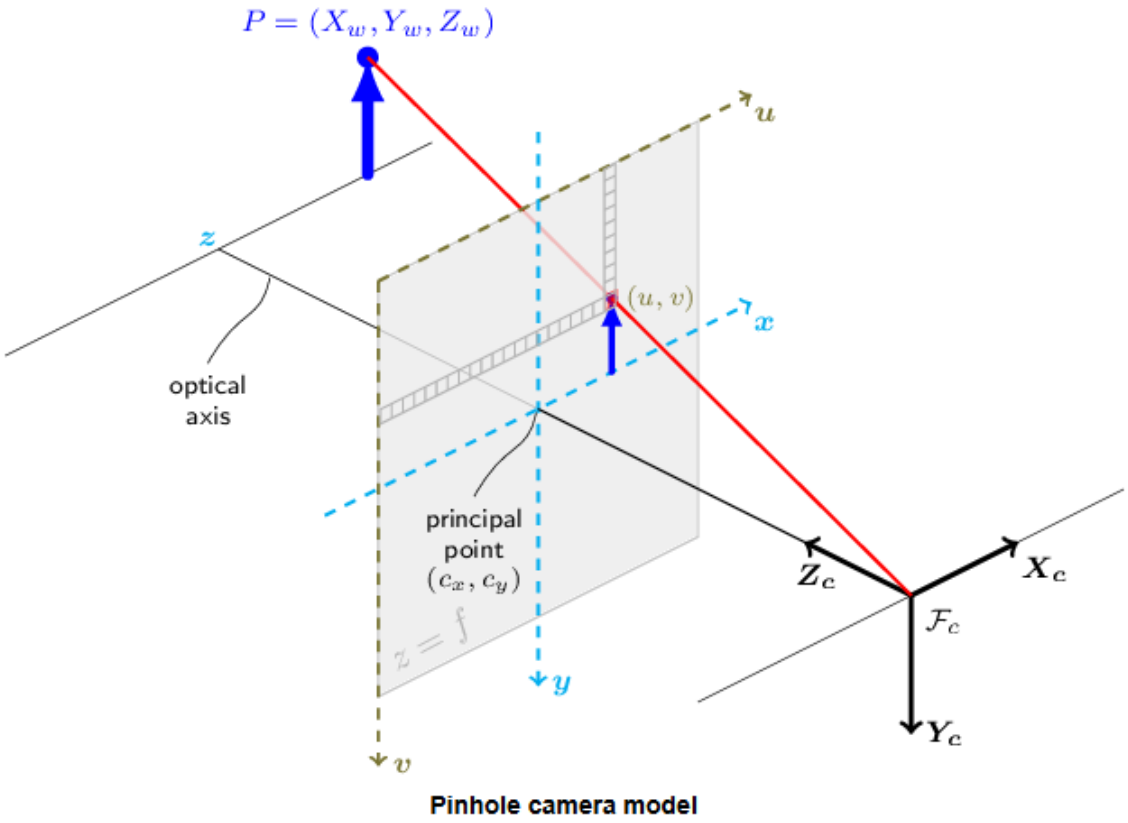

Figura 22. Representación del modelo Pinhole de transformación de puntos 3d a píxels 2d

Aunque se logró implementar el sistema de proyección de los keypoints del traje inercial sobre las imágenes capturadas, se detectaron errores sistemáticos de alineación, especialmente cuando el usuario se alejaba de la zona central de captura. En estas situaciones, algunos keypoints aparecían desplazados respecto a su posición esperada, lo que dificultaba una superposición precisa con el entorno visual. Si bien inicialmente se barajó la posibilidad de que estos desajustes se debieran a errores en la calibración de la cámara o en el vector de traslación, los análisis realizados sugieren que el origen del problema reside en la estimación de la posición absoluta por parte del propio traje Rokoko Smartsuit Pro II. Este tipo de error resulta inesperado, dado que el sistema incorpora el módulo Coil Pro como referencia global adicional para el geoposicionamiento. La aparición de esta desviación representa una limitación significativa detectada durante el proceso de creación del dataset, al impedir una alineación precisa entre los datos inerciales y las imágenes capturadas. Se prevé que futuras actualizaciones del firmware del fabricante puedan mejorar la integración del Coil Pro y solventar esta deficiencia. Cabe destacar que esta limitación ha sido también señalada por otros usuarios y miembros de la comunidad.

5.2.6. Estructuración del dataset

Se diseñó una estructura de almacenamiento coherente, modular y sistemática para organizar el conjunto de datos generado. Esta estructura permite una gestión sencilla de los archivos, facilita su uso para tareas de análisis, visualización o entrenamiento de modelos, y se inspira en formatos de datasets reconocidos internacionalmente, como el del Panoptic Studio desarrollado por la Carnegie Mellon University (CMU) (Joo et al., 2019).

El dataset resultante se compone de un total de 112 escenas únicas, correspondientes a las 14 acciones realizadas por cada uno de los 8 sujetos definidos. Para cada escena, se ha creado una carpeta con la siguiente nomenclatura:

```
sujetoX_actividadY/
```

Donde sujetoX identifica al participante y actividadY corresponde a la escena concreta que ha ejecutado. Dentro de cada una de estas 112 carpetas, se encuentran dos subdirectorios principales: imágenes y keypoints.

El subdirectorio de imágenes contiene tres subcarpetas, una por cada cámara utilizada durante la grabación:

```
sujetoX_actividadY/
│
├── imagenes/
│   ├── camara_frontal/
│   ├── camara_lateral/
│   └── camara_trasera/
```

Cada subcarpeta incluye todos los frames extraídos del vídeo original en formato .jpg, a una frecuencia uniforme de 30 fps. Las tres cámaras han sido sincronizadas mediante un dispositivo de activación simultánea, por lo que todas las subcarpetas contienen exactamente el mismo número de imágenes, con nombres correlativos (frame_0001.jpg, frame_0002.jpg, etc.).

El subdirectorio de keypoints contiene los keypoints tridimensionales estimados por el sistema Rokoko, exportados y convertidos a un formato .json por cada frame. La estructura de esta carpeta es la siguiente:

```
sujetoX_actividadY/
│
├── keypoints/
│   ├── sujetoX_actividadY_0001.json
│   ├── sujetoX_actividadY_0002.json
│   └── ...
```

Cada archivo .json contiene los datos tridimensionales correspondientes a un único frame, con una estructura basada en el formato propuesto por la Panoptic Toolbox 1. La información relevante se encuentra en el campo joints19, que contiene las coordenadas (x, y, z) y

un valor adicional asociado a la confianza o fiabilidad del punto, para un total de 19 keypoints. La estructura general es la siguiente:

```
{
  "version": 0.7,
  "univTime": 53541.542,
  "fpsType": "4k_29_97",
  "bodies": [
    {
      "id": 0,
      "joints19": [
        x0, y0, z0, c0,
        x1, y1, z1, c1,
        ...
        x18, y18, z18, c18
      ]
    }
  ]
}
```

Donde cada grupo de cuatro valores representa las coordenadas tridimensionales y la puntuación de confianza de un keypoint concreto. El orden de los keypoints aparece en la Tabla 2. No obstante, la puntuación de confianza no ha sido utilizada en este trabajo; se ha incluido únicamente por cuestiones de formato, asignando un valor constante y uniforme a todos los keypoints.

Tabla 2. Keypoints que aparecen en el dataset

Id	Nombre	Significado
0	Neck	Cuello
1	Nose	Nariz
2	BodyCenter	Centro del cuerpo
3	lShoulder	Hombro izquierdo
4	lElbow	Codo izquierdo
5	lWrist	Muñeca izquierda
6	lHip	Cadera izquierda
7	lKnee	Rodilla izquierda
8	lAnkle	Tobillo izquierdo
9	rShoulder	Hombro derecho
10	rElbow	Codo derecho
11	rWrist	Muñeca derecha
12	rHip	Cadera derecha

Id	Nombre	Significado
13	rKnee	Rodilla derecha
14	rAnkle	Tobillo derecho
15	lEye	Ojo izquierdo
16	lEar	Oreja izquierda
17	rEye	Ojo derecho
18	rEar	Oreja derecha

Este formato permite un fácil acceso a los datos para procesamiento en Python y es compatible con múltiples librerías de visión por computador y aprendizaje automático.

De forma adicional, se ha almacenado en una carpeta independiente la información técnica de las cámaras utilizadas, incluyendo los parámetros intrínsecos, coeficientes de distorsión y matrices de rotación y traslación obtenidas durante la fase de calibración. Sin embargo, tal y como se comentó en apartados anteriores, estos datos no se han utilizado de forma sistemática en las proyecciones, debido a las dificultades observadas para mantener la precisión espacial en las estimaciones del traje, especialmente cuando el sujeto se aleja del origen definido por el Coil Pro.

Por tanto, si bien la información de calibración está disponible y puede ser útil para análisis posteriores o ajustes más precisos, la estructura principal del dataset se ha definido en base a la coherencia temporal y la organización modular por sujeto y escena.

5.3. PROCESO DE EVALUACIÓN

Una vez generado el conjunto de datos estructurado y anotado en formato tridimensional, se procedió a su evaluación mediante modelos preentrenados de estimación de pose humana en 3D, entrenados previamente con grandes conjuntos de datos. El objetivo principal de esta fase es calcular la pose tridimensional a partir de las imágenes grabadas utilizando estos modelos del estado del arte, y comparar sus estimaciones con nuestro conjunto de datos desarrollado, considerándolo como *ground truth*. Esta comparación permite, por un lado, cuantificar el error de estimación y verificar la coherencia del dataset etiquetado como una forma de validación; y por otro, identificar limitaciones en los modelos existentes, especialmente en aquellos casos donde las estimaciones no resultan precisas.

Estos resultados no solo podrían evidenciar la utilidad de nuestro dataset como herramienta de evaluación y mejora, sino también, poner

de manifiesto la necesidad de nuevos conjuntos de datos etiquetados, más variados o representativos, que permitan ampliar la capacidad de generalización de estos modelos y mejorar su rendimiento en contextos más complejos o realistas.

Se seleccionaron tres modelos ampliamente utilizados en tareas de reconstrucción 3D a partir de imágenes RGB o keypoints 2D:

- MediaPipe Pose (Lugaresi et al., 2019): modelo desarrollado por Google, diseñado para funcionar en tiempo real en dispositivos móviles. Estima keypoints 2D y 3D a partir de una única imagen utilizando una arquitectura basada en BlazePose. Es especialmente ligero y eficiente.
- MHFormer (Li, 2021/2025): modelo basado en una arquitectura Transformer que toma como entrada secuencias de keypoints 2D y predice secuencias de poses 3D. Destaca por incorporar múltiples hipótesis de estimación para mejorar la robustez temporal de sus predicciones.
- MotionBERT (Zhu et al., 2023): modelo que emplea Bidirectional Encoder Representations from Transformers para capturar dependencias espaciales y temporales a partir de secuencias de keypoints 2D. Ha sido entrenado en grandes datasets de captura de movimiento y logra resultados altamente precisos.

Estos modelos se aplicaron sobre los frames del dataset experimental, extrayendo las coordenadas tridimensionales (x, y, z) de los 11 keypoints que aparecen en la Tabla 3, que son comunes tanto a los modelos evaluados como a la estructura del dataset generado. No obstante, una fortaleza adicional de este estudio es que el conjunto de datos creado contiene un mayor número de keypoints anatómicos, lo que permite entrenar modelos específicos capaces de estimar con mayor detalle otras zonas del cuerpo, ampliando así su aplicabilidad en tareas que requieren una segmentación más precisa del movimiento.

Tabla 3. Keypoints sobre los que se han aplicado los modelos con su significado

Id	Nombre	Significado
1	Nose	Nariz
4	lShoulder	Hombro izquierdo
5	lWrist	Muñeca izquierda
6	lHip	Cadera izquierda
7	lKnee	Rodilla izquierda
8	lAnkle	Tobillo izquierdo
10	rShoulder	Hombro derecho
11	rWrist	Muñeca derecha

Id	Nombre	Significado
12	rHip	Cadera derecha
13	rKnee	Rodilla derecha
14	rAnkle	Tobillo derecho

Estos keypoints se pueden visualizar en un gráfico tridimensional en la Figura 23:

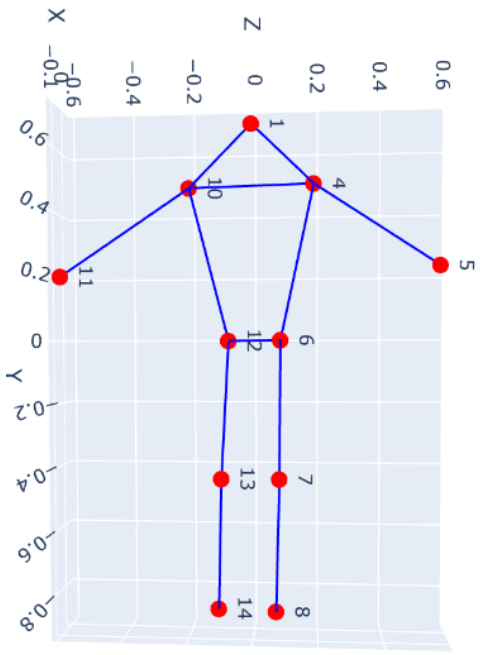

Figura 23. Visualización de los keypoints en 3 dimensiones con su número correspondiente

Para realizar una comparación justa entre las coordenadas estimadas por los modelos y los datos reales del traje, se aplicó un proceso de alineación Procrustes, cuyo objetivo es minimizar las diferencias entre dos conjuntos de puntos tridimensionales, eliminando los efectos de traslación, escala y rotación. La alineación se realizó según los siguientes pasos, basados en la metodología descrita en Procrustes analysis («Procrustes Analysis», 2025) y se puede visualizar a nivel gráfico paso por paso en la Figura 24:

1. Translación al origen: Traslación al origen: ambos conjuntos de keypoints se trasladaron para que el punto medio entre las caderas (centro de pelvis) coincidiera con el origen de coordenadas.
2. Escalado por RMSD: se normalizó cada conjunto para que la raíz del error cuadrático medio (RMSD) respecto al origen fuera 1. Dado un conjunto X', su factor de escala es:

$$s = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\|x_i'\|^2}$$

3. Rotación óptima por SSD: se determinó la matriz de rotación R que minimiza la distancia cuadrática total (SSD) entre los puntos alineados.
4. Escalado final según distancia real entre caderas: tras la alineación, se aplicó un escalado adicional para que la distancia de las caderas en los keypoints reales coincidiera con la correcta. Dicho escalado se aplicó a los keypoints del modelo para que los dos conjuntos tuvieran el mismo tamaño.

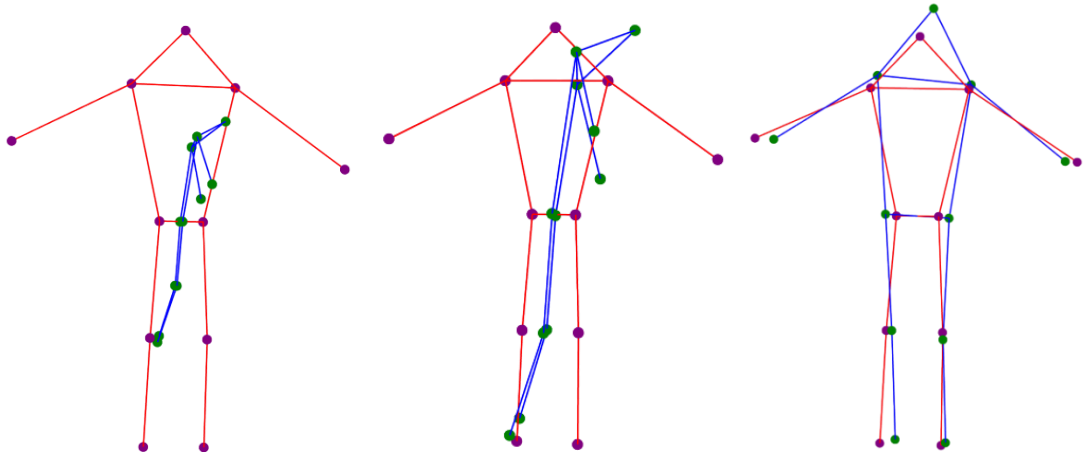

Figura 24. Representación gráfica del proceso de alineación mediante Procrustes entre la figura estimada por MediaPipe (en azul) y la figura correspondiente del conjunto de datos desarrollado (en rojo). De izquierda a derecha se muestran las tres etapas del alineamiento: translación al origen, escalado mediante la raíz del error cuadrático medio (RMSD) y rotación óptima basada en la minimización de la suma de las diferencias al cuadrado (SSD).

La métrica utilizada para evaluar el rendimiento de los modelos fue el MPJPE Procrustes o PA-MPJPE (Mean Per Joint Position Error con alineación Procrustes). Esta métrica calcula la distancia euclídea media entre cada par de puntos correspondientes tras la alineación. Formalmente, para n keypoints esta métrica se define como

$$PA - MPJPE = \frac{1}{n}\sum_{i=1}^{n}\|\hat{x}_i' - x_i\|_2,$$

donde x_i son cada keypoint del dataset creado y $\hat{x}_i'$ cada keypoint predicho por el modelo, sumando sobre los n keypoints tras la alineación Procrustes.

Como resultado de este procedimiento, para cada uno de los frames del dataset se dispone de:

- Las coordenadas tridimensionales estimadas por cada uno de los tres modelos (MediaPipe, MHFormer, MotionBERT).

- El valor de MPJPE Procrustes obtenido para cada modelo en ese frame, comparando su predicción con la referencia generada por el traje.

Esta información será utilizada en los siguientes apartados para realizar un análisis comparativo del rendimiento de los modelos, permitiendo estudiar su precisión en distintas acciones, sujetos y condiciones de captura. Además, nos servirá para evaluar nuestro dataset y detectar en qué casos específicos (ya sean acciones, sujetos o condiciones particulares) las estimaciones de los modelos difieren más, lo que permitirá extraer conclusiones sobre aquellas situaciones en las que los modelos actuales presentan limitaciones o dificultades para generalizar correctamente.

6. RESULTADOS

En este capítulo se presentan los principales resultados obtenidos en el proceso de evaluación y análisis de los modelos de 3D-HPE aplicados al conjunto de datos generado en este trabajo. La comparación se ha llevado a cabo entre tres modelos representativos del estado del arte: MediaPipe, MHFormer y MotionBERT. Para valorar su rendimiento se ha utilizado como métrica principal el error medio por articulación tras la alineación por el método de procrusters (PA-MPJPE), introducida en la sección 5.3 , tomando como referencia (valor real) las coordenadas estimadas por nuestro conjunto de datos desarrollado. Por simplicidad, en lo que sigue por MPJPE nos referiremos a la métrica PA-MPJPE, muchas veces utilizados indistintamente en la literatura.

En primer lugar, se realiza un análisis comparativo de los tres modelos evaluados. Se examina su rendimiento global mediante estadísticas descriptivas del MPJPE y se profundiza en la comparación por sujeto y tipo de acción. Este análisis aporta una visión cuantitativa y contextualizada del comportamiento de cada arquitectura.

A continuación, se abordan los outliers, es decir, aquellos casos con errores de predicción significativamente mayores en comparación con la distribución general de errores, lo que indica casos extremos donde la discrepancia entre la estimación del modelo y nuestro conjunto de datos es más pronunciada, y que requieren una revisión detallada para comprender sus causas. Se identifican estos valores atípicos más extremos por actividad y modelo, y se agrupan en función de sus características para extraer patrones comunes.

Por último, se utilizan algoritmos de aprendizaje supervisado, específicamente árboles de decisión, para analizar las causas del error de predicción. Se desarrollan modelos que clasifican las instancias según si presentan un error medio por articulación (MPJPE) alto o bajo, tomando como variables explicativas factores como la cámara empleada y la acción realizada. Este enfoque permite identificar qué condiciones o características están asociadas con un mayor error, facilitando la interpretabilidad de la causa y pudiendo extraer conclusiones fundamentadas.

De forma complementaria, en los Anexos se incluyen de manera ampliada algunas de las tablas correspondientes al análisis comparativo de modelos, tanto a nivel global como desglosado por sujeto y tipo de acción.

6.1. ANÁLISIS COMPARATIVO DE MODELOS

Se comparan tres modelos representativos en el campo de la estimación de poses humanas en 3D (MediaPipe, MHFormer y MotionBERT) con el objetivo de evaluar su precisión mediante la métrica MPJPE y analizar cómo varía su rendimiento en función de diferentes condiciones del conjunto de datos. Más allá de determinar qué modelo ofrece mejores resultados de forma global, se busca identificar en qué contextos concretos cada arquitectura se comporta mejor o peor. Este análisis es especialmente relevante para orientar la selección del modelo más adecuado en función de la aplicación: por ejemplo, en tareas biomecánicas donde se requiera precisión en movimientos dinámicos o, por el contrario, en entornos más controlados y estáticos. Asimismo, detectar debilidades sistemáticas puede resultar útil para diseñar estrategias de entrenamiento específicas que refuercen los puntos más problemáticos de cada modelo.

El análisis se plantea como una exploración abierta y empírica, sin una hipótesis previa formalizada sobre el comportamiento de los modelos. En primer lugar, se presenta una evaluación general del rendimiento de cada modelo, considerando estadísticas agregadas como la media y la mediana del MPJPE, así como otras métricas relevantes. Posteriormente, se examina cómo se ve afectado el rendimiento por la ubicación de la cámara (GoPro principal, lateral o trasera), con el fin de identificar posibles limitaciones derivadas del ángulo de captura.

En segundo lugar, se lleva a cabo un análisis desglosado por sujeto, con el objetivo de evaluar si ciertos individuos presentan sistemáticamente mayores errores de predicción. Este análisis resulta clave para determinar si los modelos son igualmente robustos frente a diferentes morfologías, estilos de movimiento o condiciones particulares del sujeto.

Por último, se presenta una comparación por tipo de actividad. Se analiza qué acciones resultan más o menos precisas (entendiéndose que difieren más o menos de nuestra estimación validada) para cada modelo y, adicionalmente, se estudia qué articulaciones del cuerpo concentran los mayores errores dentro de cada tipo de movimiento. Este enfoque aporta una perspectiva más detallada sobre las limitaciones específicas de los modelos y permite identificar patrones de fallo recurrentes asociados a ciertas posturas o movimientos.

Como se ha indicado previamente, a partir de este punto consideraremos como "más preciso" o de "menor error" aquellos casos en los que la estimación de los modelos difiera en menor medida respecto a la estimación validada de nuestro conjunto de datos desarrollado.

6.1.1. Evaluación general

En este apartado se analiza el rendimiento general de los tres modelos evaluados a partir del valor del MPJPE. El objetivo es comparar

la precisión global de cada arquitectura y estudiar su comportamiento según el punto de vista de la imagen, determinado por las tres cámaras GoPro empleadas: una frontal (GP3), una lateral (GP2) y una trasera (GP4).

Como se puede observar en la Tabla 4 el modelo MHFormer obtiene los mejores resultados en cuanto a media, seguido por MotionBERT y, en último lugar, MediaPipe. Esta tendencia se confirma también en la mediana, los cuartiles y la desviación típica, lo que sugiere una mayor estabilidad del modelo MHFormer frente a los otros dos. Los valores de MPJPE de MotionBERT son en general algo mejores que los de MediaPipe, aunque la diferencia es menos acusada.

Tabla 4. Análisis descriptivo del valor de MPJPE para cada uno de los modelos

Estadístico	Media (DE)	Mediana (IQR)	Max	Outliers
Mediapipe	0.1192 (0.0457)	0.1089 (0.0896 - 0.1366)	0.6921	14010
MHFormer	0.0963 (0.0374)	0.0880 (0.0724 - 0.1096)	0.4509	14820
MotionBERT	0.1148 (0.0427)	0.1050 (0.0875 - 0.1293)	0.5776	18449

El análisis gráfico mediante diagramas de caja representado en la Figura 25 muestra que MHFormer presenta menor dispersión y menor número de valores atípicos extremos, lo que refuerza su consistencia. MediaPipe, por el contrario, aunque muestra menor cantidad de outliers, cuando estos aparecen, su magnitud es considerablemente mayor. MotionBERT destaca por tener un número elevado de valores extremos, aunque estos no alcanzan la magnitud de los observados en MediaPipe.

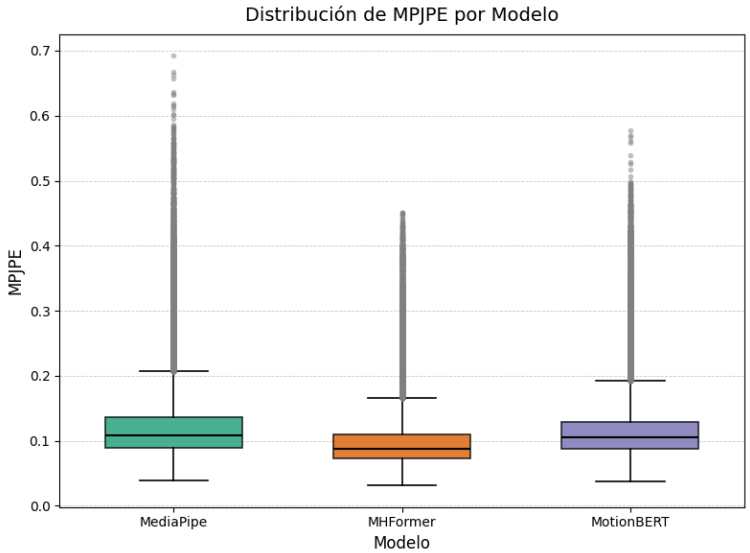

Figura 25. Boxplot de los valores de MPJPE para cada uno de los modelos

Por otro lado, si se analiza el rendimiento según la vista de la cámara, se aprecia en la Figura 26 que la cámara frontal (GP3) ofrece sistemáticamente los mejores valores de MPJPE en los tres modelos, tanto en media como en mediana. Esto sugiere que una vista frontal

favorece una estimación más precisa de la postura, como era de esperar, ya que presenta menos dificultades relacionadas con la oclusión y ofrece una mejor visibilidad de las articulaciones clave. Además, esta perspectiva facilita una mayor consistencia en la detección de puntos clave, lo que contribuye a reducir el error en la estimación.

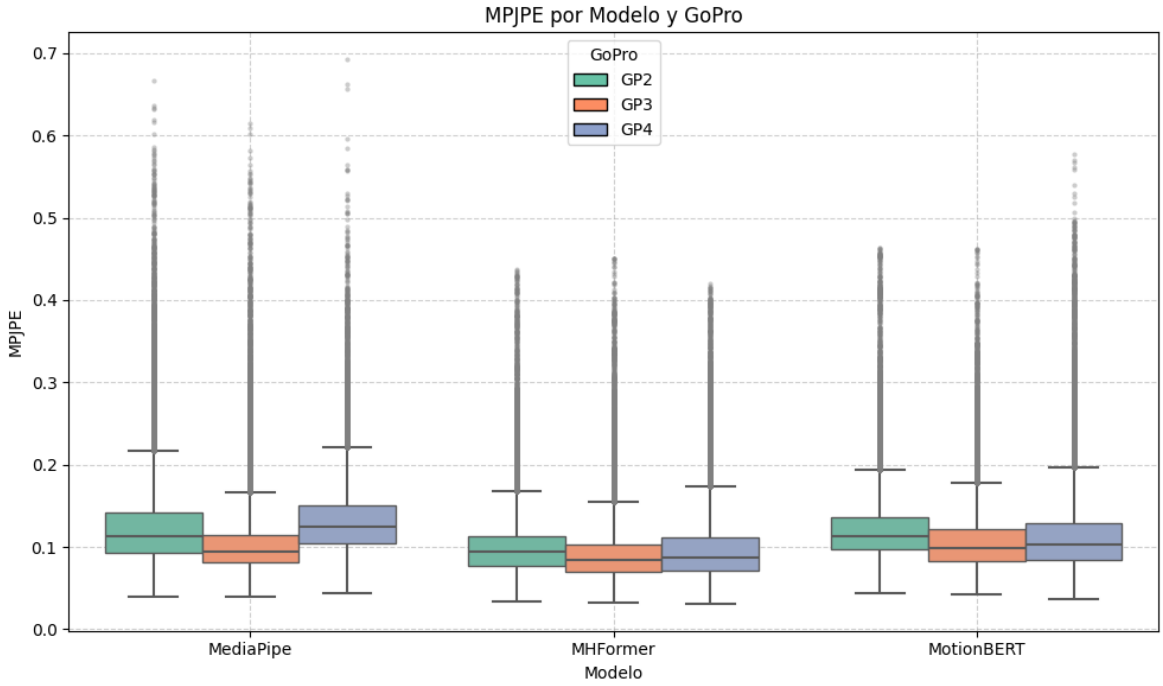

Figura 26. Boxplot del valor de MPJPE en cada una de las vistas, para cada uno de los modelos

En cuanto a la comparación entre GP2 (lateral) y GP4 (trasera), se observan diferencias relevantes entre modelos. En MHFormer y MotionBERT, la GP4 ofrece mejores medias y medianas que la GP2, mientras que en MediaPipe ocurre lo contrario: la vista lateral presenta mejores resultados que la trasera.

En los tres puntos de vista (localizaciones de las cámaras), MHFormer (véase Tabla 5) obtiene sistemáticamente los mejores valores de media, mediana y cuartiles, lo que refuerza su posición como el modelo más preciso y estable tanto en diferentes actividades como en diferentes perspectivas. En GP2 y GP3, MediaPipe (véase Tabla 6) ofrece resultados ligeramente mejores en los cuartiles inferiores (menos error en las predicciones más acertadas), mientras que MotionBERT (véase Tabla 7) presenta ligeramente mejores valores en los cuartiles superiores, lo que indica que tiende a ser más robusto ante errores moderados. En la GP4, MotionBERT supera a MediaPipe en todos los estadísticos considerados.

Tabla 5. Análisis descriptivo del valor de MPJPE para MHFormer

GoPro	Media (DE)	Mediana (IQR)	Max	Outliers
GP2	0.0999 (0.0349)	0.0936 (0.0776 - 0.1134)	0.4372	3,611
GP3	0.0920 (0.0354)	0.0838 (0.0696 - 0.1034)	0.4509	6,092

GoPro	Media (DE)	Mediana (IQR)	Max	Outliers
GP4	0.0978 (0.0412)	0.0870 (0.0712 - 0.1122)	0.4195	6,388

Tabla 6. Análisis descriptivo del valor de MPJPE para MediaPipe

GoPro	Media (DE)	Mediana (IQR)	Max	Outliers
GP2	0.1247 (0.0507)	0.1124 (0.0926 - 0.1425)	0.6671	4,077
GP3	0.1037 (0.0389)	0.0949 (0.0812 - 0.1151)	0.6149	3,611
GP4	0.1325 (0.0431)	0.1241 (0.1043 - 0.1508)	0.6921	4,418

Tabla 7. Análisis descriptivo del valor de MPJPE para MotionBERT

GoPro	Media (DE)	Mediana (IQR)	Max	Outliers
GP2	0.1214 (0.0385)	0.1137 (0.0973 - 0.1360)	0.4628	4,418
GP3	0.1080 (0.0379)	0.0986 (0.0830 - 0.1210)	0.4622	6,388
GP4	0.1159 (0.0502)	0.1025 (0.0847 - 0.1292)	0.5776	7,643

6.1.2. Comparación por sujeto

En el siguiente apartado se analiza el rendimiento de los modelos en función del sujeto que realiza la acción. Este enfoque permite evaluar si existen diferencias sistemáticas de error entre los sujetos seleccionados en el apartado 5.1.5, así como identificar posibles patrones relacionados con la morfología, la ropa, el estilo de movimiento u otros factores particulares que puedan influir en la estimación de pose.

En los tres modelos analizados, se observa una tendencia clara, observable en la Figura 27 y la Figura 28, donde se presentan las distribuciones del error MPJPE por sujeto, para cada uno de los modelos. Los sujetos 6 y 2 presentan valores ligeramente más bajos de MPJPE, tanto en términos de media como de mediana y cuartiles. Por el contrario, el sujeto 4 registra el mayor error medio, seguido por el sujeto 7. Esta jerarquía en el rendimiento se mantiene consistente en las tres arquitecturas evaluadas: MediaPipe (Tabla 8), MHFormer (Tabla 9) y MotionBERT (Tabla 10). No obstante, las diferencias son muy ligeras y podría considerarse que se comportan para todos sujetos de manera más o menos similar en términos generales

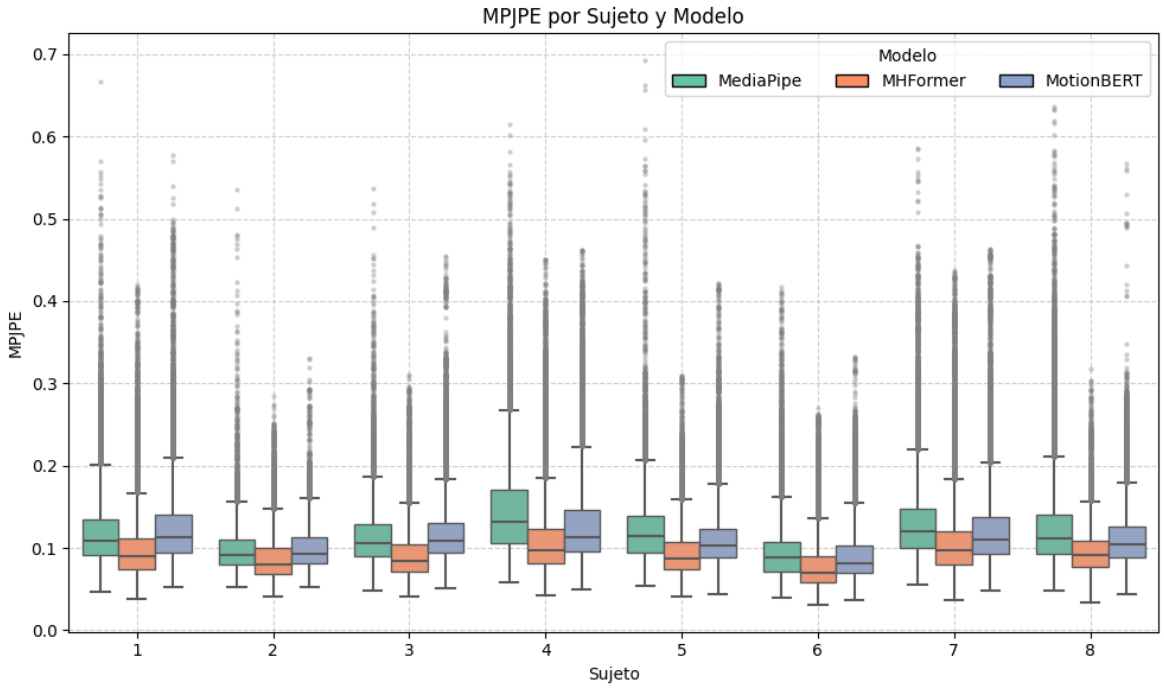

Figura 27. Boxplot del valor de MPJPE para cada uno de los modelos, en cada uno de los sujetos

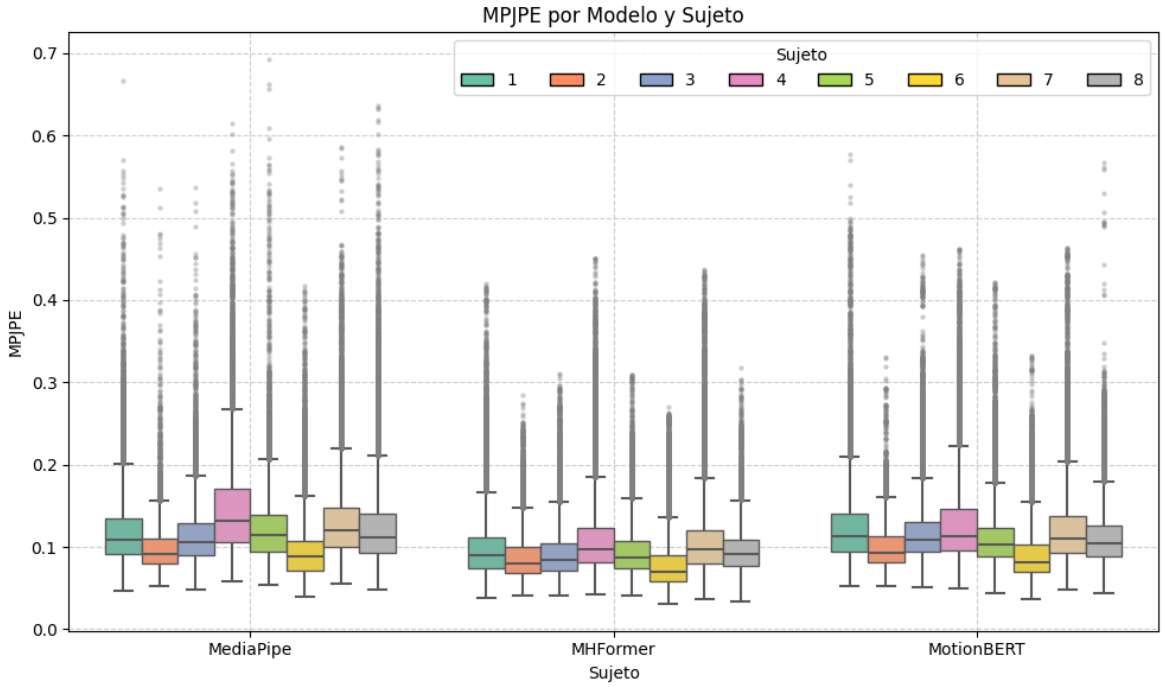

Figura 28. Boxplot del valor de MPJPE en cada uno de los sujetos, para cada uno de los modelos

Tabla 8. Análisis descriptivo de los valores de MPJPE en cada uno de los sujetos para el modelo MediaPipe

Sujeto	Nº	Media (DE)	Mediana (IQR)	Min	Max	% outliers
1	51448	0.119 (0.042)	0.109 (0.135 - 0.091)	0.046	0.667	4.09
2	14306	0.099 (0.031)	0.092 (0.111 - 0.080)	0.052	0.535	1.15
3	42073	0.113 (0.034)	0.106 (0.128 - 0.089)	0.048	0.537	1.69
4	28169	0.145 (0.057)	0.132 (0.170 - 0.105)	0.058	0.615	9.69

Sujeto	Nº	Media (DE)	Mediana (IQR)	Min	Max	% outliers
5	39099	0.122 (0.040)	0.115 (0.140 - 0.095)	0.054	0.692	2.87
6	40920	0.095 (0.036)	0.088 (0.108 - 0.072)	0.039	0.417	1.53
7	43985	0.130 (0.045)	0.121 (0.148 - 0.100)	0.055	0.586	5.17
8	33074	0.126 (0.056)	0.111 (0.140 - 0.092)	0.048	0.637	6.35

Tabla 9. Análisis descriptivo de los valores de MPJPE en cada uno de los sujetos para el modelo MHFormer

Sujeto	Nº	Media (DE)	Mediana (IQR)	Min	Max	% outliers
1	51909	0.099 (0.037)	0.090 (0.111 - 0.074)	0.038	0.419	5.23
2	14378	0.088 (0.030)	0.081 (0.100 - 0.068)	0.041	0.284	3.07
3	42318	0.092 (0.030)	0.084 (0.105 - 0.071)	0.041	0.310	2.98
4	35730	0.111 (0.049)	0.098 (0.123 - 0.081)	0.042	0.451	11.01
5	28133	0.095 (0.031)	0.087 (0.108 - 0.074)	0.040	0.310	3.63
6	42273	0.078 (0.030)	0.070 (0.090 - 0.058)	0.031	0.270	1.97
7	44280	0.107 (0.043)	0.097 (0.121 - 0.079)	0.036	0.437	7.26
8	41127	0.096 (0.028)	0.091 (0.109 - 0.077)	0.033	0.317	2.65

Tabla 10. Análisis descriptivo de los valores de MPJPE en cada uno de los sujetos para el modelo MotionBERT

Sujeto	Nº	Media (DE)	Mediana (IQR)	Min	Max	% outliers
1	51909	0.124 (0.047)	0.113 (0.140 - 0.094)	0.053	0.578	8.04
2	14378	0.102 (0.031)	0.092 (0.113 - 0.081)	0.053	0.331	1.16
3	42318	0.117 (0.038)	0.109 (0.130 - 0.095)	0.051	0.455	4.86
4	35730	0.129 (0.053)	0.113 (0.147 - 0.096)	0.049	0.462	11.30
5	39333	0.111 (0.036)	0.103 (0.124 - 0.088)	0.043	0.422	3.43
6	42273	0.091 (0.032)	0.082 (0.103 - 0.069)	0.037	0.332	1.50
7	44280	0.122 (0.047)	0.110 (0.137 - 0.092)	0.049	0.463	7.45
8	41127	0.112 (0.034)	0.104 (0.125 - 0.089)	0.044	0.568	3.45

El porcentaje de outliers por sujeto refuerza estas observaciones. El sujeto 4 es el que acumula mayor proporción de outliers en los tres modelos, con un valor superior al 9,7 % en MediaPipe, 11 % en MHFormer y 11,3 % en MotionBERT. Por el contrario, los sujetos 6 y 2 presentan porcentajes muy bajos de valores extremos, lo que confirma que los modelos son menos propensos a errores ante estos sujetos.

El análisis desglosado por cámara ofrece información adicional relevante, como se muestra en la Tabla 11. En MediaPipe, el sujeto 4 es el que más outliers acumula en las tres cámaras, mientras que el sujeto 8 destaca en la GP2. En MHFormer, el sujeto 4 vuelve a ser el más problemático, seguido del sujeto 7 en la GP4. En MotionBERT, el comportamiento se repite, con el sujeto 4 como el más conflictivo, seguido por el 1 y el 7, especialmente en la cámara trasera.

Tabla 11. Porcentaje de valores de MPJPE que son outlier en cada sujeto correspondientes a cada GoPro (GP2, GP3 y GP4)), para cada modelo

	MediaPipe			MHFormer			MotionBERT		
Sujeto	GP2	GP3	GP4	GP2	GP3	GP4	GP2	GP3	GP4
1	4.98	1.70	5.58	4.71	2.42	8.58	6.17	6.35	11.60
3	1.55	0.82	2.70	3.69	0.96	4.28	6.94	0.82	6.83
4	8.60	7.80	12.67	9.50	12.83	10.71	9.11	12.02	12.75
5	1.71	1.50	5.38	1.95	2.73	6.04	2.82	1.61	5.87
6	0.39	2.20	1.99	0.59	4.09	1.23	0.33	2.55	1.61
7	4.46	2.35	8.69	4.92	6.58	10.29	4.40	4.74	13.22
8	14.14	1.48	3.52	2.81	1.98	3.17	3.45	1.71	5.17

Si bien los resultados muestran que existen diferencias en el error de estimación entre sujetos concretos, la variabilidad observada podría estar asociada a condiciones externas de grabación como la iluminación, la ropa utilizada, la orientación frente a las cámaras o pequeños desajustes en la sincronización. Por tanto, no se puede atribuir con certeza dicha sensibilidad únicamente a las características individuales de los sujetos, sino que es posible que actúen como variables confusoras otras condiciones contextuales del entorno de captura.

6.1.3. Comparación por tipo de acción

A continuación, se analiza el rendimiento de los modelos MediaPipe, MHFormer y MotionBERT en función del tipo de acción realizada por el sujeto, cuyas características están en el apartado 5.1.6. Este análisis permite identificar qué actividades resultan más complejas para cada arquitectura y si existen patrones de error asociados a dinámicas corporales concretas.

Los boxplots correspondientes a la Figura 29 y a la Figura 30 muestran el valor de MPJPE para cada uno de los modelos en cada una de las acciones. MHFormer (véase Tabla 13) obtiene los mejores resultados en todas las acciones, tanto en media como en mediana del MPJPE. MotionBERT (véase Tabla 14) se posiciona como el segundo con mejores resultados de media y mediana, superando ligeramente a MediaPipe (véase Tabla 12) en la mayoría de las actividades, con excepción de las acciones 5 (vestirse) y 13 (sentarse), donde MediaPipe presenta un menor error medio.

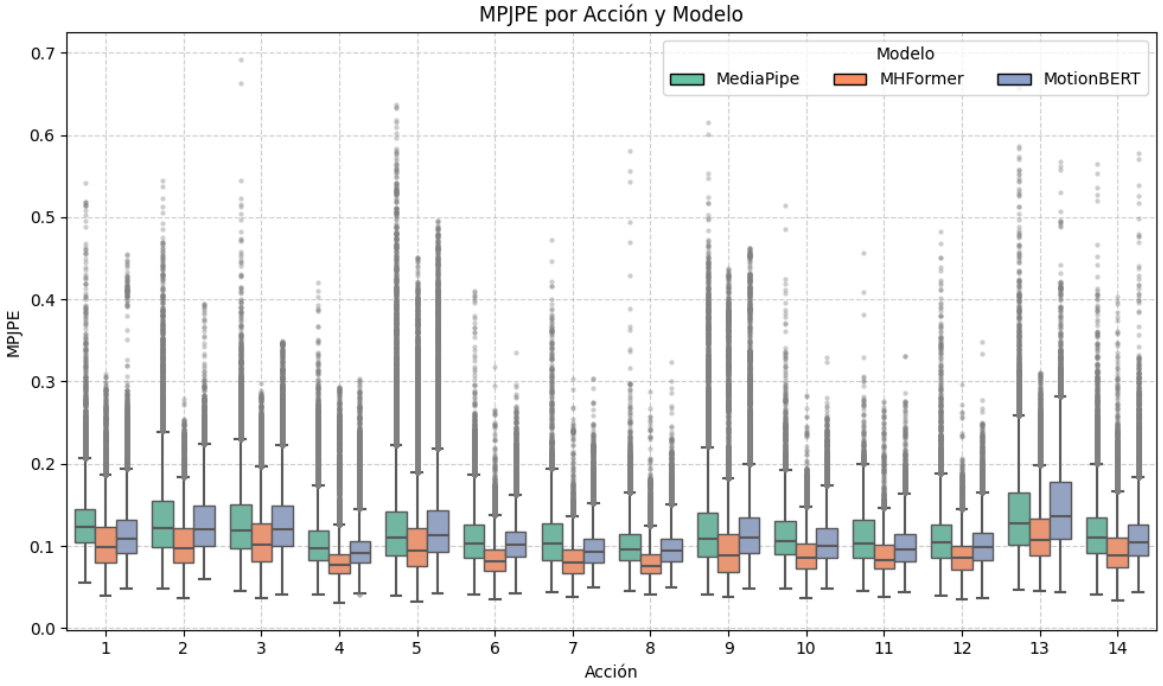

Figura 29. Boxplot del valor de MPJPE para cada uno de los modelos, en cada una de las acciones

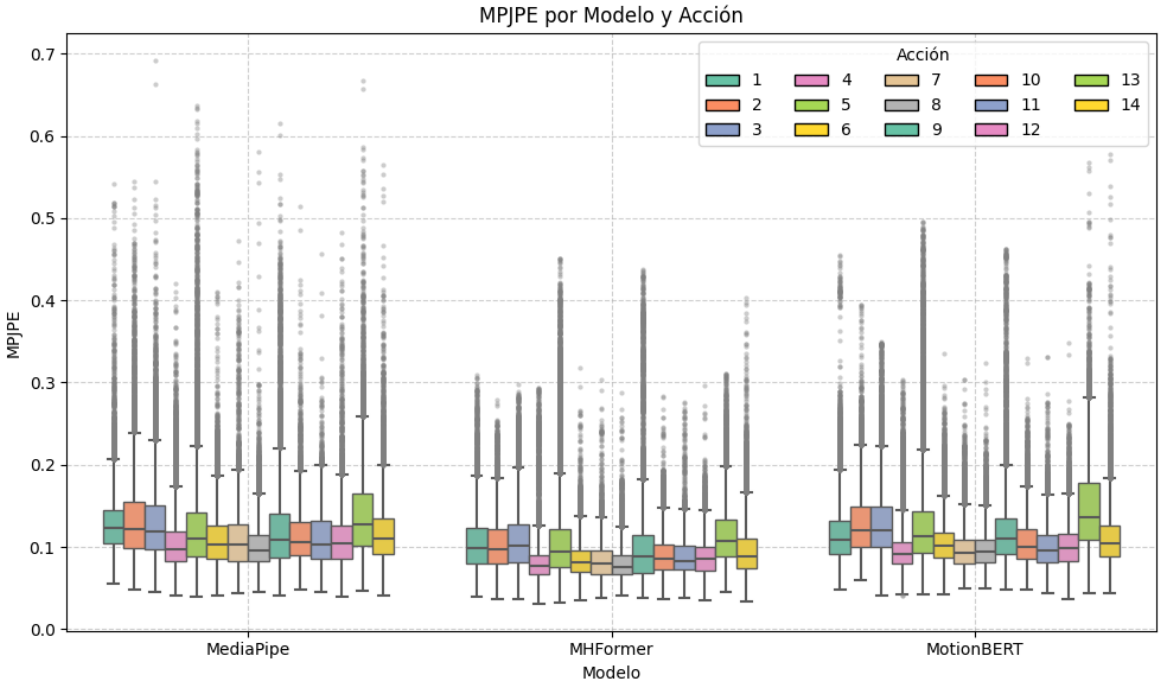

Figura 30. Boxplot del valor de MPJPE en cada una de las acciones, para cada uno de los modelos

Tabla 12. Análisis descriptivo de los valores de MPJPE en cada una de las acciones para el modelo MediaPipe

Acción	Media (DE)	Mediana (IQR)	Max	% outliers
1	0.130 (0.038)	0.124 (0.105 - 0.145)	0.541	3.14
2	0.134 (0.055)	0.121 (0.098 - 0.155)	0.545	8.35
3	0.131 (0.049)	0.119 (0.098 - 0.151)	0.692	7.21

Acción	Media (DE)	Mediana (IQR)	Max	% outliers
4	0.107 (0.037)	0.097 (0.083 - 0.119)	0.421	3.14
5	0.125 (0.060)	0.110 (0.089 - 0.142)	0.637	7.02
6	0.110 (0.035)	0.103 (0.086 - 0.126)	0.410	1.80
7	0.108 (0.036)	0.103 (0.082 - 0.127)	0.472	1.11
8	0.102 (0.030)	0.096 (0.082 - 0.115)	0.581	0.62
9	0.123 (0.057)	0.109 (0.087 - 0.141)	0.615	6.66
10	0.114 (0.034)	0.106 (0.090 - 0.131)	0.515	1.21
11	0.111 (0.034)	0.103 (0.086 - 0.131)	0.456	0.77
12	0.110 (0.037)	0.104 (0.086 - 0.127)	0.482	1.59
13	0.139 (0.053)	0.128 (0.102 - 0.165)	0.667	10.30
14	0.118 (0.040)	0.110 (0.091 - 0.135)	0.565	3.17

Tabla 13. Análisis descriptivo de los valores de MPJPE en cada una de las acciones para el modelo MHFormer

Acción	Media (DE)	Mediana (IQR)	Max	% outliers
1	0.105 (0.035)	0.099 (0.081 - 0.123)	0.309	5.50
2	0.103 (0.032)	0.098 (0.080 - 0.121)	0.278	3.80
3	0.109 (0.038)	0.101 (0.082 - 0.128)	0.298	8.45
4	0.084 (0.034)	0.076 (0.066 - 0.090)	0.293	4.17
5	0.109 (0.054)	0.095 (0.076 - 0.121)	0.451	11.32
6	0.084 (0.022)	0.082 (0.069 - 0.096)	0.317	0.47
7	0.083 (0.022)	0.079 (0.067 - 0.095)	0.303	0.35
8	0.080 (0.021)	0.076 (0.066 - 0.089)	0.287	0.47
9	0.103 (0.055)	0.089 (0.068 - 0.114)	0.437	9.50
10	0.090 (0.026)	0.086 (0.072 - 0.103)	0.284	1.71
11	0.089 (0.025)	0.083 (0.072 - 0.102)	0.276	0.84
12	0.088 (0.024)	0.086 (0.071 - 0.101)	0.297	0.71
13	0.115 (0.038)	0.107 (0.089 - 0.133)	0.311	10.08
14	0.097 (0.034)	0.089 (0.074 - 0.111)	0.403	4.53

Tabla 14. Análisis descriptivo de los valores de MPJPE en cada una de las acciones para el modelo MotionBERT

Acción	Media (DE)	Mediana (IQR)	Max	% outliers
1	0.117 (0.040)	0.109 (0.091 - 0.132)	0.455	5.14
2	0.128 (0.040)	0.121 (0.100 - 0.149)	0.395	7.72
3	0.129 (0.043)	0.120 (0.100 - 0.149)	0.349	8.41
4	0.099 (0.032)	0.092 (0.081 - 0.107)	0.303	3.24
5	0.129 (0.059)	0.114 (0.093 - 0.144)	0.496	11.46
6	0.104 (0.026)	0.101 (0.087 - 0.117)	0.335	1.11
7	0.098 (0.025)	0.093 (0.081 - 0.109)	0.304	0.83
8	0.098 (0.024)	0.094 (0.081 - 0.109)	0.323	0.49
9	0.121 (0.051)	0.111 (0.091 - 0.134)	0.463	6.82
10	0.107 (0.030)	0.100 (0.086 - 0.121)	0.329	1.73
11	0.101 (0.028)	0.096 (0.081 - 0.114)	0.332	0.82
12	0.102 (0.028)	0.099 (0.083 - 0.116)	0.348	0.81
13	0.149 (0.056)	0.136 (0.109 - 0.178)	0.568	19.54
14	0.112 (0.039)	0.105 (0.088 - 0.127)	0.578	3.83

La distribución del error es coherente entre los tres modelos: las acciones con mayor error son generalmente aquellas que implican movimientos amplios, cambios posturales o interacción con objetos, como 2 (agacharse), 3 (estiramientos), 5 (vestirse), 9 (foto) y 13 (sentarse). Por el contrario, las acciones con menor error son aquellas que implican movimientos más estáticos o repetitivos, como 7 (teléfono), 8 (fumar), 11 (música) y 12 (bailar).

El análisis del porcentaje de outliers por acción refuerza estos hallazgos. En MediaPipe, las actividades con mayor proporción de valores atípicos son 2, 3, 5, 9 y 13, mientras que las menos conflictivas son 8 (fumar) y 11 (música). En MHFormer, se observa una mayor incidencia de outliers en 1 (caminar), 3 (estiramientos), 5 (vestirse), 9 (foto) y 13 (sentarse). En MotionBERT, destaca notablemente la actividad 13 (sentarse) como la más problemática, seguida por 5 (vestirse). Las actividades más estables en este modelo son 7 (teléfono), 8 (fumar), 11 (música) y 12 (bailar).

Al analizar los resultados segregados por cámara, presentados en la Tabla 15, se observan patrones consistentes con las observaciones anteriores. En MediaPipe, las cámaras laterales (GP2) y trasera (GP4) presentan mayores dificultades en la acción 13 (sentarse), mientras que en la cámara principal (GP3) el error se distribuye de forma más equilibrada, con cierta mayor incidencia en las acciones 2 (agacharse), 3 (estiramientos) y 5 (vestirse). En MHFormer, la cámara trasera es especialmente sensible en la acción 5 (vestirse), que presenta un porcentaje de outliers significativamente superior al resto. La cámara lateral muestra más dificultades en 1 (caminar), 3 (estiramientos) y 9 (foto), mientras que la frontal presenta los mayores valores en 3 (estiramientos), 9 (foto) y 13 (sentarse). Por su parte, en MotionBERT, la cámara trasera vuelve a mostrar un comportamiento crítico en 13 (sentarse), con un porcentaje de outliers muy elevado. También destacan 5 (vestirse) y 2 (agacharse) como acciones problemáticas en esta vista.

Tabla 15. Porcentaje de valores de MPJPE que son outlier en cada acción correspondientes a cada GoPro (GP2, GP3 y GP4)), para cada modelo

	MediaPipe			MHFormer			MotionBERT		
Acción	GP2	GP3	GP4	GP2	GP3	GP4	GP2	GP3	GP4
1	4.68	1.98	2.97	7.08	4.99	4.52	6.87	3.67	5.13
2	8.30	5.42	11.79	1.82	2.58	7.18	2.77	9.60	10.51
3	7.74	5.10	9.04	7.46	9.83	7.89	9.03	8.50	7.69
4	3.10	2.35	4.07	4.24	3.96	4.34	5.11	1.02	3.94
5	6.44	5.51	9.35	6.21	5.12	23.56	6.40	6.13	22.66
6	1.80	0.10	3.72	0.56	0.36	0.51	1.43	1.63	0.20
7	1.79	0.42	1.21	0.15	0.59	0.26	1.06	1.30	0.07
8	1.03	0.13	0.78	0.22	1.01	0.08	1.07	0.35	0.08

	MediaPipe			MHFormer			MotionBERT		
9	8.57	4.41	7.30	10.26	8.33	10.09	5.99	7.83	6.51
10	2.25	0.51	0.98	2.61	1.13	1.49	2.04	1.94	1.17
11	0.89	0.17	1.32	0.11	0.29	2.20	0.79	0.36	1.37
12	4.20	0.23	0.52	0.95	0.45	0.76	1.42	0.72	0.30
13	11.30	2.52	18.49	4.74	13.23	10.95	12.45	6.74	41.73
14	5.85	0.49	3.56	6.56	2.09	5.30	5.46	1.29	5.09

Estos resultados muestran cómo el tipo de acción y el punto de vista de la cámara influyen directamente en la precisión de los modelos. Las tareas que implican posturas complejas u oclusiones corporales tienden a generar más error, especialmente en ciertas combinaciones de modelo y cámara. Esto es coherente, ya que cuando partes del cuerpo están ocultas o las posturas implican movimientos complejos, asimetrías o cambios bruscos de orientación, los modelos tienen más dificultades para hacer estimaciones precisas, especialmente si no han sido entrenados con ejemplos que reflejen esas condiciones. Esta observación pone en evidencia y refuerza la necesidad sobre la que se basa nuestro estudio: desarrollar un conjunto de datos con actividades diversas y realistas que permita mejorar la robustez de los modelos actuales y ampliar su capacidad de generalización.

6.1.3.1. Comparación por articulaciones

Este subapartado recoge un análisis detallado del error cometido por los modelos en función de la articulación corporal estimada. Las articulaciones existentes (y su identificador) se pueden observar en el apartado 5.2.6. El objetivo es identificar qué regiones anatómicas presentan mayores dificultades para cada modelo y cómo varía esta tendencia según el tipo de acción realizada.

Considerando todos los modelos en conjunto, en la Tabla 16 se muestran los 4 keypoints con mayor error para cada acción. Se observa que las muñecas (derecha e izquierda) son sistemáticamente las articulaciones con mayor error de estimación en prácticamente todas las actividades. En segundo lugar, el error tiende a concentrarse en los tobillos o en la cabeza, dependiendo del tipo de acción. Concretamente, en acciones más estáticas como hablar, fumar o dar direcciones, es la cabeza la que suele presentar un mayor error, mientras que en actividades que implican más movimiento o desplazamiento (como caminar, agacharse, estirarse o sentarse), el error se incrementa notablemente en los tobillos. Estos hallazgos coinciden con lo que cabía esperar.

Tabla 16. Keypoints con mayor error (por orden), para todos los modelos en conjunto

accion	Kp1	Errorkp1	Kp2	Errorkp2	Kp3	Errorkp3	Kp4	Errorkp4
1	5	0.18	11	0.175	8	0.165	14	0.164
2	5	0.199	11	0.18	14	0.138	1	0.134
3	11	0.208	5	0.205	8	0.148	14	0.143
4	5	0.156	11	0.149	1	0.117	14	0.112
5	5	0.203	11	0.196	1	0.136	14	0.135
6	11	0.192	5	0.184	1	0.119	8	0.105
7	11	0.175	5	0.166	14	0.109	1	0.106
8	11	0.17	5	0.166	1	0.104	14	0.1
9	5	0.209	11	0.2	1	0.134	14	0.131
10	5	0.201	11	0.196	1	0.117	14	0.11
11	11	0.191	5	0.188	14	0.117	8	0.112
12	11	0.185	5	0.172	8	0.113	1	0.109
13	5	0.207	11	0.205	8	0.168	14	0.165
14	11	0.202	5	0.189	14	0.122	8	0.12

Al analizar individualmente cada modelo como en la Tabla 17, se mantienen en general las mismas tendencias, aunque con matices específicos:

- En MediaPipe, las muñecas vuelven a ser las articulaciones con mayor error en la mayoría de actividades. En acciones estáticas, les sigue la cabeza, mientras que en las más dinámicas son los tobillos los que presentan un mayor error. Cabe destacar que el error en la cabeza es más pronunciado que en los otros modelos, lo que podría estar relacionado con diferencias en el punto anatómico que el modelo proyecta como referencia (por ejemplo, la nariz en lugar del centro de la cabeza).
- En MHFormer, las muñecas también encabezan el ranking de error. A diferencia de MediaPipe, el segundo lugar lo ocupan mayoritariamente los tobillos, y solo en muy pocas acciones la cabeza supera en error a alguna de las extremidades inferiores. Esto sugiere una mayor consistencia en la estimación del segmento superior del cuerpo respecto a MediaPipe.
- En MotionBERT, se repite la predominancia de error en las muñecas, seguidas de los tobillos. Sin embargo, en este modelo también aparecen ocasionalmente valores elevados en la cabeza y, en algunos casos, en los hombros, lo que indica una mayor variabilidad en el reparto del error entre diferentes articulaciones.

Tabla 17. Keypoints (Kp1, Kp2, Kp3 y Kp4) con mayor error en cada acción, para cada modelo

		MediaPipe		MHFormer		MotionBERT	
Acción	Top-4	KP	MPJPE	KP	MPJPE	KP	MPJPE
1	Kp1	5	0.225	5	0.153	11	0.165

		MediaPipe		MHFormer		MotionBERT	
	Kp2	11	0.211	11	0.151	5	0.164
	Kp3	8	0.18	8	0.151	8	0.164
	Kp4	14	0.179	14	0.15	14	0.164
2	Kp1	5	0.241	5	0.166	5	0.192
	Kp2	11	0.215	11	0.148	11	0.179
	Kp3	14	0.162	14	0.12	1	0.14
	Kp4	8	0.152	8	0.119	7	0.136
3	Kp1	5	0.238	11	0.185	11	0.206
	Kp2	11	0.234	5	0.182	5	0.196
	Kp3	8	0.161	8	0.135	8	0.149
	Kp4	14	0.16	14	0.129	14	0.142
4	Kp1	5	0.189	5	0.134	11	0.153
	Kp2	1	0.167	11	0.127	5	0.149
	Kp3	11	0.167	8	0.094	14	0.115
	Kp4	14	0.135	1	0.088	8	0.109
5	Kp1	5	0.235	5	0.181	5	0.195
	Kp2	11	0.218	11	0.178	11	0.193
	Kp3	14	0.15	1	0.124	1	0.145
	Kp4	1	0.141	8	0.123	8	0.141
6	Kp1	5	0.218	11	0.164	11	0.203
	Kp2	11	0.21	5	0.153	5	0.182
	Kp3	1	0.165	8	0.09	1	0.114
	Kp4	8	0.13	1	0.08	10	0.107
7	Kp1	11	0.215	11	0.156	11	0.158
	Kp2	5	0.201	5	0.143	5	0.156
	Kp3	1	0.142	14	0.086	14	0.112
	Kp4	8	0.131	8	0.085	10	0.104
8	Kp1	5	0.203	11	0.154	11	0.17
	Kp2	11	0.187	5	0.139	5	0.159
	Kp3	1	0.146	8	0.082	10	0.11
	Kp4	14	0.12	14	0.08	14	0.102
9	Kp1	5	0.232	5	0.189	5	0.206
	Kp2	11	0.212	11	0.181	11	0.206
	Kp3	1	0.166	8	0.118	1	0.131
	Kp4	14	0.156	14	0.112	14	0.124
10	Kp1	5	0.227	11	0.182	5	0.194
	Kp2	11	0.217	5	0.178	11	0.189
	Kp3	1	0.16	8	0.095	10	0.112
	Kp4	14	0.129	14	0.09	14	0.108
11	Kp1	5	0.224	11	0.179	11	0.176
	Kp2	11	0.218	5	0.174	5	0.166
	Kp3	14	0.137	14	0.098	14	0.114
	Kp4	1	0.136	8	0.095	10	0.113

		MediaPipe		MHFormer		MotionBERT	
12	Kp1	11	0.209	11	0.167	11	0.179
	Kp2	5	0.207	5	0.147	5	0.161
	Kp3	1	0.139	8	0.095	8	0.106
	Kp4	8	0.137	14	0.092	14	0.105
13	Kp1	5	0.227	11	0.18	11	0.218
	Kp2	11	0.215	5	0.178	5	0.214
	Kp3	8	0.188	14	0.15	8	0.177
	Kp4	14	0.181	8	0.139	14	0.165
14	Kp1	11	0.221	11	0.184	11	0.201
	Kp2	5	0.219	5	0.166	5	0.182
	Kp3	1	0.149	8	0.107	14	0.12
	Kp4	14	0.141	14	0.105	1	0.115

Estos resultados reflejan la sensibilidad de los modelos a articulaciones más lejanas al origen, que tienden a ser más difíciles de estimar con precisión, especialmente en condiciones de oclusión, movimiento acelerado o perspectivas desfavorables. Este comportamiento es coherente con lo esperado, ya que estas articulaciones presentan mayor variabilidad y menor visibilidad en comparación con partes más centrales del cuerpo.

6.2. ANÁLISIS DE OUTLIERS

En el siguiente apartado se analizan los outliers extremos identificados en cada actividad para los tres modelos evaluados. Se ha considerado el outlier del frame con el mayor valor de MPJPE por actividad y modelo, con el fin de identificar los errores más significativos y detectar posibles patrones comunes entre ellos.

6.2.1. Outliers en Mediapipe

Los outliers más graves en MediaPipe presentan errores notablemente altos, con valores de MPJPE que en algunos casos superan los 0.6. En muchas actividades, el sujeto más afectado es el sujeto 4. Este sujeto utiliza una vestimenta de camiseta multicolor con mangas negras, lo cual podría dificultar la segmentación y detección correcta de las extremidades superiores. También se han observado errores destacados en el sujeto 1, cuya ropa ancha podría contribuir a errores de estimación. Los outliers pueden clasificarse en diferentes categorías, según su posible origen:

- **Outliers de vestimenta**, como se muestra en la Figura 31, están vinculados a prendas que dificultan la identificación de los

contornos corporales. Esta categoría incluye especialmente al sujeto 4 y, en menor medida, al sujeto 1.

- **Outliers de acción**, como a modo de ejemplo se ilustra en la Figura 32 (actividad de agacharse), donde se observa un mayor error en actividades con posturas extremas, como agacharse (actividad 2 y 3) o sentarse en una silla (actividad 13).
- **Outliers de oclusión**, visibles en la Figura 33, ocurren cuando partes del cuerpo están fuera del encuadre o parcialmente tapadas, tal y como ocurre en la figura. Por ejemplo, en algunos frames el sujeto aparece cortado por el borde de la imagen o excesivamente cerca de la cámara.
- **Outliers por condiciones de iluminación**, recogidos en la Figura 34, se producen cuando factores como contraluz o sombra dificultan la visibilidad de extremidades, especialmente si coinciden con ropa oscura. Esto ocurre en el caso del sujeto 8 en la vista lateral, donde los brazos se camuflan con el fondo, generando un error elevado.

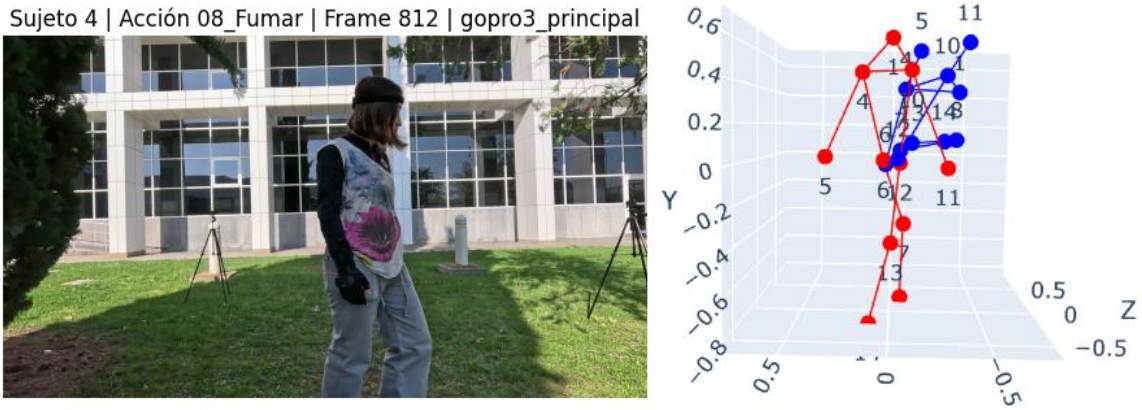

Figura 31. Comparativa entre un fotograma (izquierda) y su representación tridimensional (derecha). En rojo se muestra la posición real capturada por el traje de captura de movimiento, mientras que en azul se representa la predicción estimada por el modelo MediaPipe

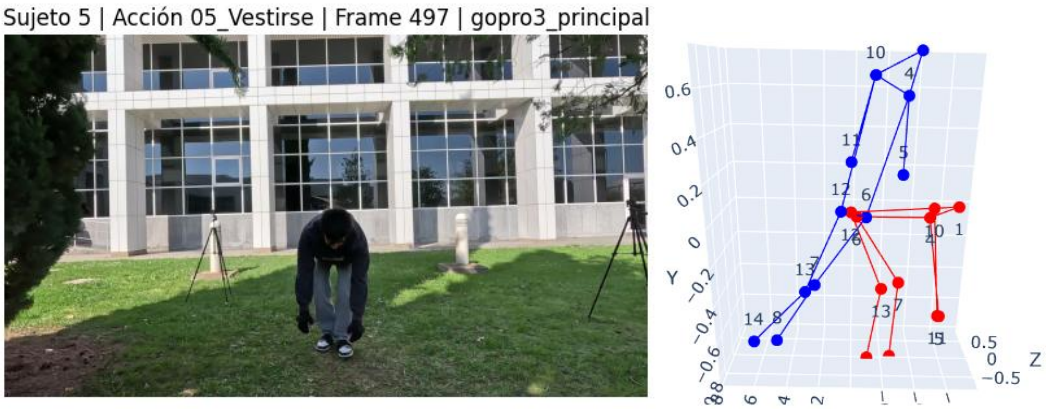

Figura 32. Comparativa entre un fotograma (izquierda) y su representación tridimensional (derecha). En rojo se muestra la posición real capturada por el traje de captura de movimiento, mientras que en azul se representa la predicción estimada por el modelo MediaPipe.

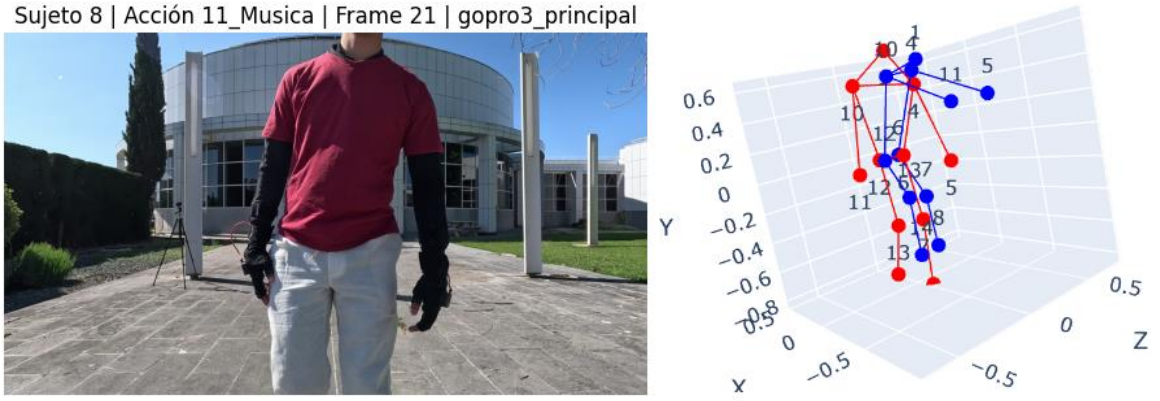

Figura 33. Comparativa entre un fotograma (izquierda) y su representación tridimensional (derecha). En rojo se muestra la posición real capturada por el traje de captura de movimiento, mientras que en azul se representa la predicción estimada por el modelo MediaPipe.

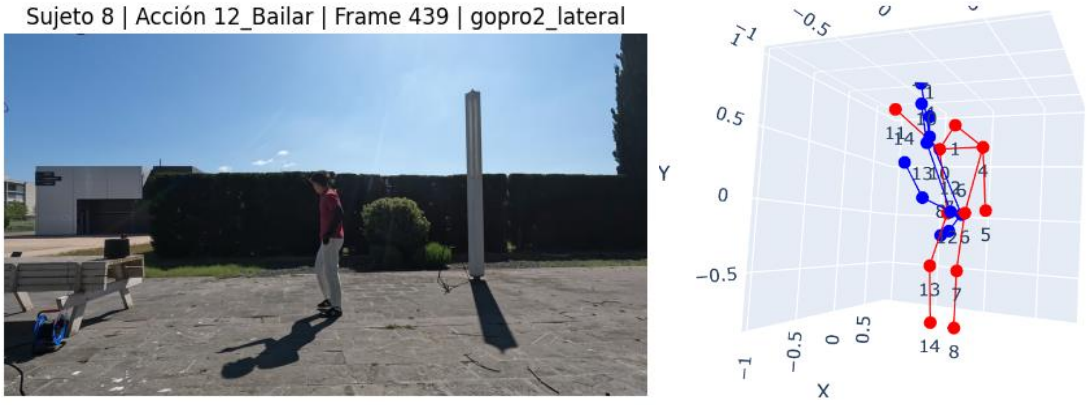

Figura 34. Comparativa entre un fotograma (izquierda) y su representación tridimensional (derecha). En rojo se muestra la posición real capturada por el traje de captura de movimiento, mientras que en azul se representa la predicción estimada por el modelo MediaPipe.

Cabe destacar el elevado rango de error de los outliers en este modelo, con varios casos por encima de 0.6 MPJPE, lo que evidencia una menor robustez frente a condiciones adversas.

6.2.2. Outliers en MHFormer

En el caso de MHFormer, los valores de MPJPE en los outliers más altos son considerablemente menores que en MediaPipe, sin superar los 0.45 en ningún caso. Los frames con mayor error se concentran principalmente en acciones con posturas atípicas, como agacharse (actividades 2, 3 y 5) y sentarse (actividad 13). También se han identificado outliers causados por oclusiones, como sujetos cortados en los bordes del encuadre (actividad 14) o presencia de personas de fondo detectadas erróneamente (actividad 11). Los outliers pueden clasificarse en dos grandes grupos:

- **Outliers de acción**, como se ilustra en la Figura 35 (actividad de sentarse), relacionados con posturas como agacharse o sentarse, aunque con errores moderados.

- **Outliers de oclusión**, representados en la Figura 36, donde el sujeto aparece parcialmente fuera del encuadre o se produce confusión con personas del fondo.

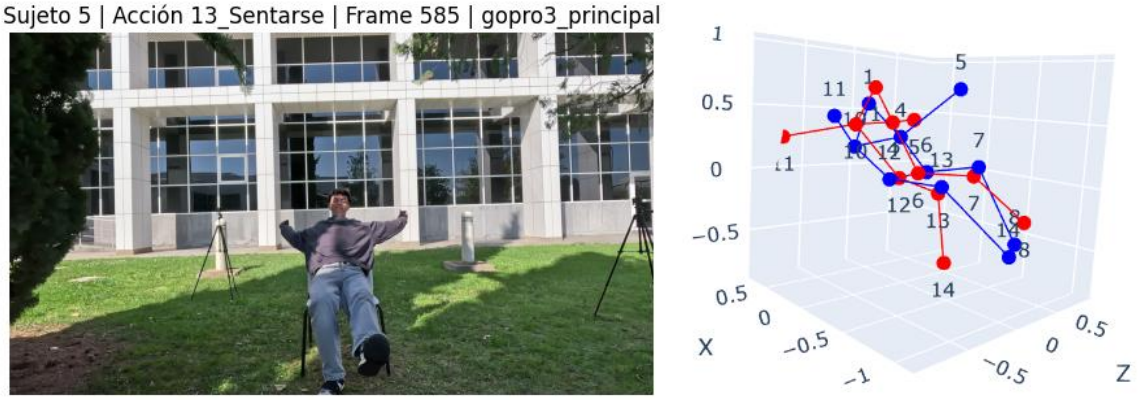

Figura 35. Comparativa entre un fotograma (izquierda) y su representación tridimensional (derecha). En rojo se muestra la posición real capturada por el traje de captura de movimiento, mientras que en azul se representa la predicción estimada por el modelo MHFormer

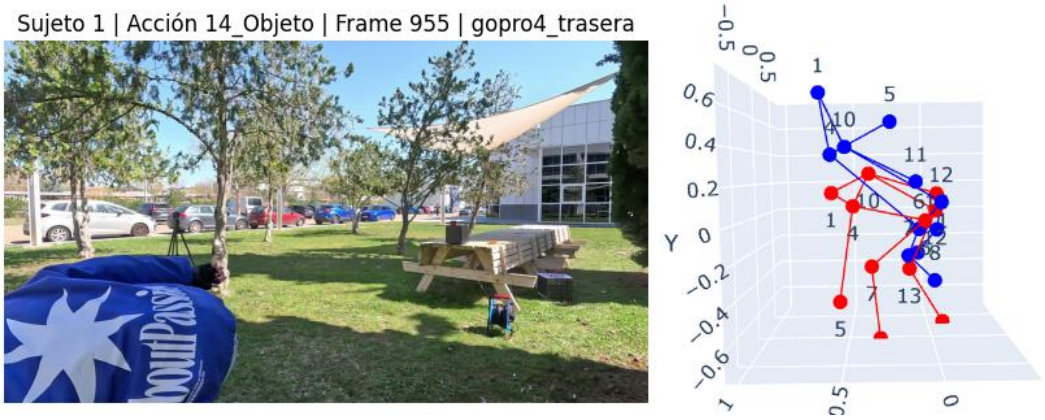

Figura 36. Comparativa entre un fotograma (izquierda) y su representación tridimensional (derecha). En rojo se muestra la posición real capturada por el traje de captura de movimiento, mientras que en azul se representa la predicción estimada por el modelo MHFormer

Pese a estas situaciones, MHFormer mantiene una estabilidad mucho mayor, con errores controlados incluso en los frames más problemáticos.

Aunque los outliers por oclusión son en gran medida esperables debido a limitaciones inherentes a los modelos basados en visión, otros tipos de error como los asociados a determinadas acciones (por ejemplo, posturas extremas como agacharse o sentarse) reflejan limitaciones del modelo entrenado con datos estándar. Este tipo de error podría reducirse mediante un reentrenamiento del modelo con conjuntos de datos que incluyan dichas variaciones posturales, como el propuesto en este trabajo. De este modo, se evidencia la utilidad de contar con datasets más diversos y específicos que ayuden a mejorar el desempeño de los modelos en situaciones menos convencionales.

6.2.3. Outliers en MotionBERT

MotionBERT presenta un comportamiento intermedio. Aunque sus outliers tienen errores menores que los de MediaPipe, son ligeramente superiores a los observados en MHFormer. Las situaciones que provocan mayor error coinciden en gran parte con las de los modelos anteriores: acciones con postura agachada (actividades 2, 3 y 5), sentado en silla (actividad 13) y oclusiones severas (actividad 14). También se ha detectado un caso de detección errónea de una persona de fondo (actividad 11). Los tipos de outliers identificados son:

- **Outliers de acción**, tal como se representa en la Figura 37 (actividad de agacharse), aparecen en actividades con posturas no estándar (agachado o sentado).
- **Outliers de oclusión**, como se muestra en la Figura 38, aparecen especialmente en frames donde el sujeto está parcialmente fuera del encuadre.

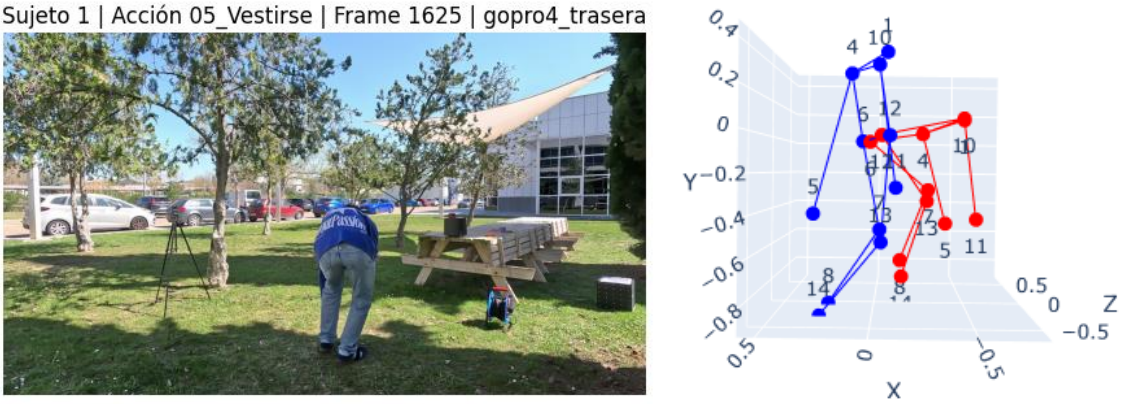

Figura 37. Comparativa entre un fotograma (izquierda) y su representación tridimensional (derecha). En rojo se muestra la posición real capturada por el traje de captura de movimiento, mientras que en azul se representa la predicción estimada por el modelo MotionBERT

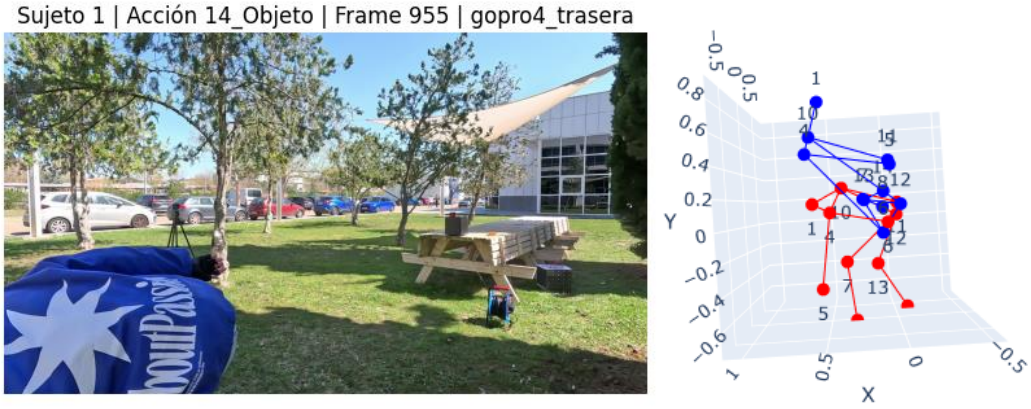

Figura 38. Comparativa entre un fotograma (izquierda) y su representación tridimensional (derecha). En rojo se muestra la posición real capturada por el traje de captura de movimiento, mientras que en azul se representa la predicción estimada por el modelo MotionBERT

En resumen, mientras que MediaPipe presenta los outliers más extremos y variados en sus causas (ropa, posturas, iluminación, oclusión), MHFormer se muestra como el modelo más robusto, con errores limitados incluso en situaciones adversas. MotionBERT, por su

parte, muestra un rendimiento más consistente que MediaPipe, aunque con cierta sensibilidad a condiciones similares.

6.3. EXPLICABILIDAD MEDIANTE ALGORITMOS DE APRENDIZAJE SUPERVISADO

Con el objetivo de analizar la influencia relativa de las variables cámara y actividad sobre el error de estimación, se ha aplicado un algoritmo de clasificación supervisada mediante árboles de decisión *DecisionTreeClassifier*. El problema se ha planteado como una tarea de clasificación binaria distinguiendo entre valores (errores) altos y bajos (según el valor del MPJPE), considerando como punto de corte la mediana, tras aplicar one-hot encoding a las variables categóricas correspondientes a la posición de la cámara y a la acción realizada. Se ha evaluado tanto un modelo general considerando el error conjunto obtenido a partir de los tres modelos de estimación de pose, como modelos específicos construidos por separado para analizar el comportamiento individual de cada uno. La métrica de evaluación considerada fue la accuracy, con el objetivo de medir la proporción de aciertos en la clasificación del error como alto o bajo.

En el árbol de decisión que integra el conjunto de los resultados de MediaPipe, MHFormer y MotionBERT mostrado en la Figura 39, alcanza un accuracy de clasificación del 63 %. La variable con mayor poder discriminativo es la GP3 (cámara frontal), situada en el nodo raíz del árbol. El decision tree asocia esta gopro a valores bajos de MPJPE, lo que es coherente ya que es la vista en la que se ve el cuerpo con menor oclusión. A partir de este punto, las acciones que más se relacionan con valores altos de error son, por orden de importancia, la actividad 13 (sentarse), 3 (estiramientos), 1 (caminar) y 2 (agacharse), mientras que las que tienden a generar menores errores son las actividades 4 (hablar), 8 (fumar) y 7 (teléfono). Estos resultados confirman un patró claro: las acciones con mayor dinamismo corporal tienden a generar más error, mientras que las acciones estáticas favorecen una estimación más precisa.

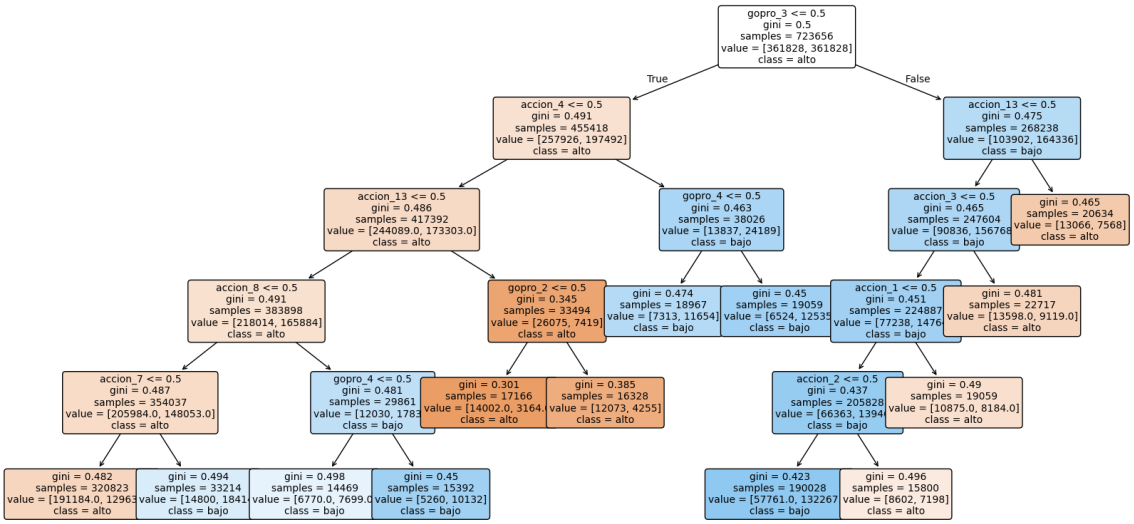

Figura 39. Árbol de decisión utilizado para clasificar valores altos y bajos del error MPJPE para los tres modelos en conjunto, en función de la cámara GoPro utilizada y la acción realizada.

Aunque el modelo se ha evaluado sobre el conjunto completo de errores estimados, este enfoque ha permitido identificar patrones e indicios relevantes sobre el comportamiento individual de cada modelo, lo que ha facilitado la extracción de algunas conclusiones. A pesar de que el rendimiento del clasificador no es especialmente alto, las variables analizadas (la posición de la cámara y la actividad realizada) revelan tendencias consistentes que ponen de manifiesto ciertas limitaciones en los modelos de estimación de pose. Sin embargo, analizar cada modelo por separado puede permitir capturar mejor sus patrones de error, ya que las diferencias entre modelos pueden "diluir" el rendimiento del análisis conjunto. Bajo esta hipótesis, se entrenó un modelo de árbol de decisión para cada modelo.

En el caso del modelo MediaPipe, el árbol de decisión entrenado se presenta en la Figura 40, el modelo alcanza un accuracy del 67 %, ligeramente superior al modelo combinado. Esto podría sugerir que las variables consideradas (cámara y actividad) permiten explicar con mayor claridad el comportamiento de sus errores de estimación en el caso de MediaPipe. La variable con mayor peso sigue siendo la GP3 (cámara frontal), cuya presencia se asocia a un menor error. En la segunda jerarquía del árbol aparece la GP4 (cámara trasera), cuya presencia se asocia, por el contrario, a mayores valores de MPJPE. En cuanto a las acciones, destacan con error alto la actividad 1 (caminar) y la actividad 13 (sentarse), mientras que las acciones con menor error son la actividad 4 (hablar) y la actividad 6 (dar direcciones). Esto sugiere que MediaPipe presenta más dificultades en contextos de movimiento dinámico y en vistas traseras, mientras que es más preciso en condiciones estáticas y con vista frontal.

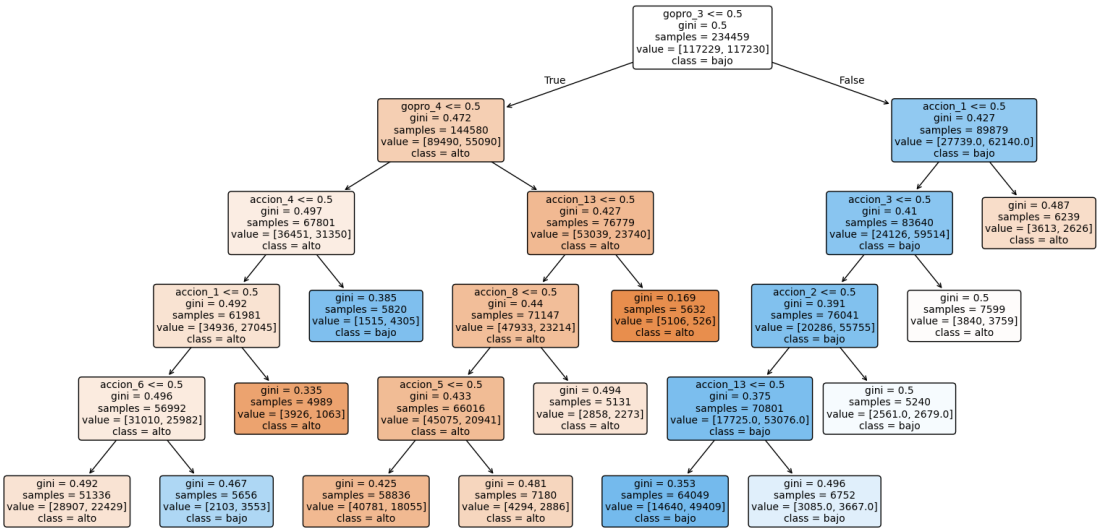

Figura 40. Árbol de decisión utilizado para clasificar valores altos y bajos del error MPJPE en el modelo MediaPipe, en función de la cámara GoPro utilizada y la acción realizada.

Para MHFormer, el árbol de decisión entrenado se presenta en la Figura 41, con un accuracy del 63 %. A diferencia de los anteriores, la variable con mayor poder de clasificación no es la cámara, sino la actividad 13 (sentarse), situada como nodo raíz. Su presencia se asocia consistentemente a valores altos de MPJPE. También se observa que la GP2 (cámara lateral) presenta una ligera tendencia a errores más elevados en diferentes acciones. En cuanto al resto de actividades, las que contribuyen a errores altos (en menor medida que la actividad 13) son la actividad 3 (estiramientos), la 1 (caminar) y la 5 (vestirse), mientras que las que más se asocian a errores bajos son la actividad 4 (hablar), 8 (fumar) y 7 (teléfono). El patrón vuelve a ser coherente: acciones dinámicas generan más error, mientras que acciones estáticas permiten una mejor estimación.

En este modelo, a diferencia de MediaPipe, la posición de la cámara no parece tener un impacto tan determinante en el error de estimación que la actividad (en concreto, sentarse).

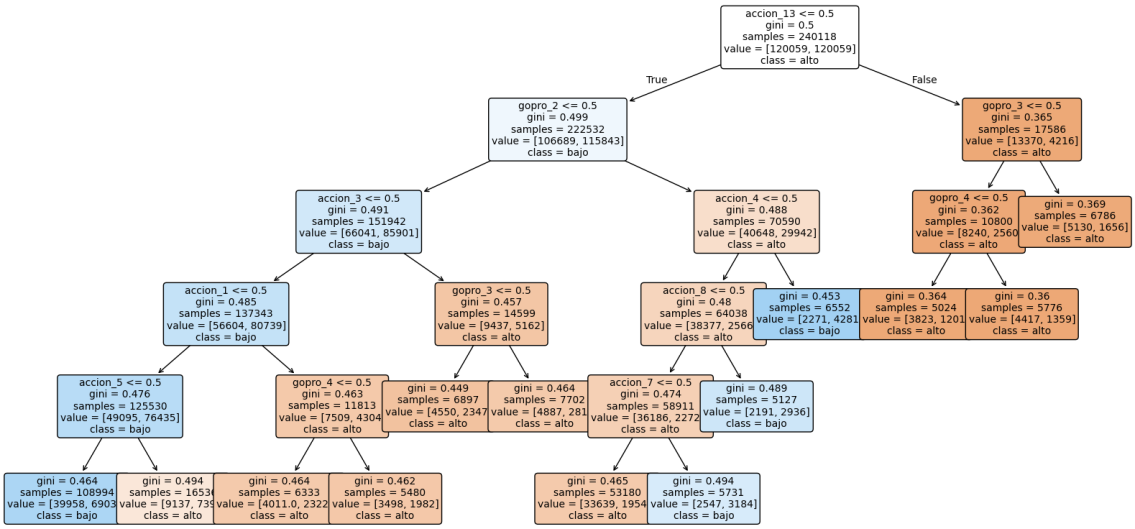

Figura 41. Árbol de decisión utilizado para clasificar valores altos y bajos del error MPJPE en el modelo MHFormer, en función de la cámara GoPro utilizada y la acción realizada.

En el modelo MotionBERT, el árbol de decisión entrenado aparece en la Figura 42 y tiene un accuracy ligeramente inferior, del 61 %. En este caso, la variable más influyente es la GP2, situada en el nodo raíz: su presencia se asocia sistemáticamente a un mayor error de estimación. Entre las actividades con mayor error destacan la actividad 13 (sentarse) y la actividad 3 (estiramientos); en cambio, las de menor error son de nuevo la actividad 4 (hablar), 8 (fumar) y 7 (teléfono). Este modelo muestra una sensibilidad particular a las imágenes tomadas desde una vista lateral, siendo este el ángulo que más contribuye al aumento del error.

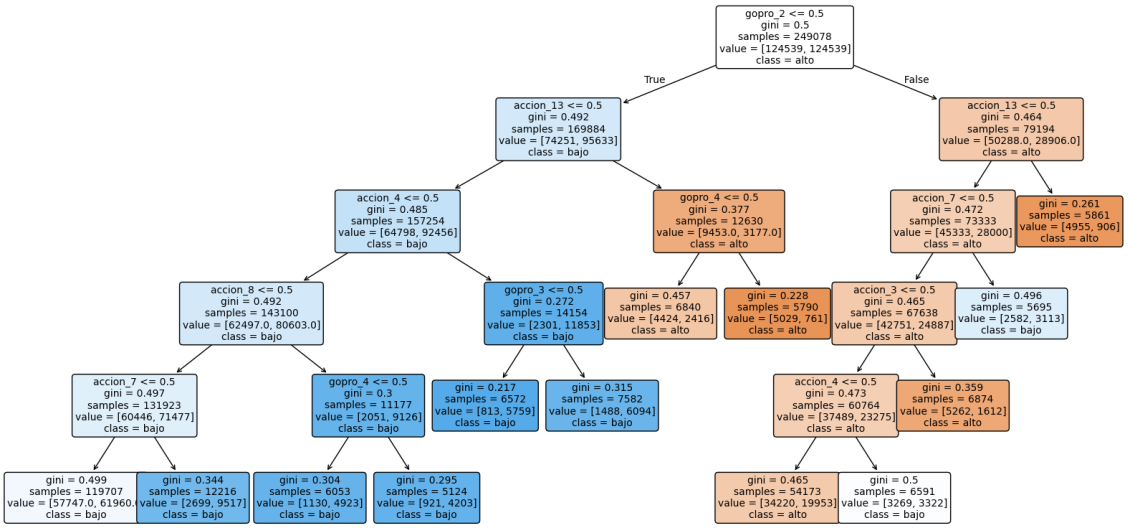

Figura 42. Árbol de decisión utilizado para clasificar valores altos y bajos del error MPJPE en el modelo MotoionBERT, en función de la cámara GoPro utilizada y la acción realizada.

En conclusión, el error cometido por MediaPipe y MotionBERT se asocia principalmente con la cámara, especialmente en el caso de

MediaPipe, donde la posición de la cámara tiene un impacto más notable en la precisión. Sin embargo, para MHFormer, la influencia de la cámara es menos determinante y el error se relaciona más directamente con el tipo de actividad realizada, en concreto la de sentarse. Esto indica que, aunque la posición de la cámara puede afectar la calidad de la estimación, las características específicas de las actividades y sus dinámicas también juegan un papel crucial en el desempeño de los modelos, resaltando la necesidad de incluir una amplia variedad de movimientos en los conjuntos de datos para mejorar su robustez.

7. DISCUSIÓN Y CONCLUSIONES

Este Trabajo de Fin de Grado se ha planteado con el objetivo de abordar una limitación relevante en el campo de la estimación de la pose humana en 3D: la escasez de conjuntos de datos públicos que integren capturas de movimiento mediante sensores inerciales con anotaciones precisas de keypoints en 3D, sincronizadas con imágenes en entornos no controlados. Frente a los métodos tradicionales, que requieren infraestructuras costosas y ambientes controlados (como en el caso de datasts como Human3.6M o MPI-INF-3DHP), se ha propuesto una metodología alternativa basada en el uso del traje inercial Rokoko Smartsuit Pro II y un sistema de grabación multicámara.

El principal objetivo ha sido la generación de un conjunto de datos propio, destinado a la creación y validación de modelos de inteligencia artificial. El proceso de construcción del dataset ha implicado un trabajo exhaustivo, que ha abarcado desde la preparación del entorno de grabación, la exploración y aplicación de estrategias de calibración, la sincronización de las diferentes fuentes de información (Rokoko y tres cámaras), hasta la extracción, sincronización y proyección de los keypoints obtenidos a partir de las imágenes.

Dicho conjunto se ha creado siguiendo la metodología CRISP-DM, y ha permitido la evaluación de tres modelos representativos (Mediapipe, MHFormer y MotionBERT), empleando tanto métricas cuantitativas (MPJPE) como análisis cualitativos mediante técnicas de explicabilidad.

La evaluación realizada con modelos del estado del arte, ampliamente utilizados y validados en tareas de estimación de pose, ha permitido una doble valoración: por un lado, validar la calidad, coherencia y utilidad del conjunto de datos desarrollado; y por otro, evidenciar que en determinadas actividades y condiciones, estos modelos presentan un mayor error, lo que pone de manifiesto la necesidad de datasets más diversos, específicos y representativos para mejorar su rendimiento, como el desarrollado en este TFG.

Los modelos analizados presentan diferencias estructurales relevantes. MediaPipe realiza una estimación directa en 3D a partir de imágenes RGB, lo que le permite operar en tiempo real con bajo coste computacional, aunque su rendimiento se ve afectado por oclusiones, cercanía a la cámara o posturas complejas. Por su parte, MHFormer y MotionBERT utilizan un enfoque de lifting, primero estimando keypoints en 2D y luego proyectándolos a 3D, lo que favorece una mayor precisión. MHFormer destaca por su solidez ante cambios de vista o tipo de acción, gracias a su capacidad para capturar relaciones espaciales y temporales. MotionBERT también logra resultados consistentes, aunque muestra algo más de variabilidad en función del contexto.

En términos globales, MHFormer ha demostrado ser el modelo más preciso y estable (considerando como más preciso aquel que se encuentra más próximo a la estimación obtenida en el conjunto de datos desarrollado), con un rendimiento uniforme en las tres vistas de cámara (frontal, lateral y trasera). Su capacidad para mantener valores consistentes, independientemente de la orientación de la cámara, sugiere una generalización efectiva ante variaciones geométricas. Por su parte, MotionBERT ha logrado una cobertura superior de fotogramas, aunque con una tasa más elevada de outliers, especialmente en la vista trasera. Mediapipe, si bien genera menos outliers, presenta errores más significativos en escenas con poses complejas o en condiciones de proximidad con la cámara.

El análisis por sujetos revela una variabilidad significativa en el error de estimación, lo que sugiere que factores individuales no controlados (como la vestimenta, el ajuste del traje inercial, la iluminación o la complexión física) podrían influir en el desempeño. No obstante, al no haber sido estas variables objeto directo del estudio, no es posible determinar con certeza su impacto específico. Se ha observado que el sujeto 6 presenta sistemáticamente mejores resultados, posiblemente debido a unas condiciones de grabación más favorables y una sincronización más precisa. En cambio, el sujeto 4 ha presentado dificultades recurrentes con Mediapipe, probablemente asociadas a la ausencia de mangas, lo cual dificulta la segmentación de extremidades.

En relación con las acciones, aquellas que implican un mayor dinamismo, como agacharse, caminar, estirarse o sentarse, generan una mayor cantidad de errores y outliers. Por el contrario, las acciones más estáticas, como hablar o fumar, presentan un mejor desempeño. Las actividades 5 (vestirse) y 13 (sentarse) han resultado especialmente problemáticas para todos los modelos, si bien MHFormer ha mitigado en parte estos errores gracias a su capacidad de modelado temporal.

Respecto al análisis por articulaciones, las muñecas (izquierda y derecha) han sido identificadas como los puntos más conflictivos, seguidas de los tobillos y la cabeza. Esta dificultad se asocia a la alta movilidad de estas articulaciones, su lejanía del centro del cuerpo y su limitada visibilidad en numerosas escenas. En actividades estáticas, se ha registrado mayor error en la cabeza, mientras que en acciones dinámicas el error se ha trasladado a los tobillos. Estas variaciones responden tanto a factores geométricos como a limitaciones en la articulación de zonas distales por parte de los modelos.

El análisis de explicabilidad mediante árboles de decisión ha permitido identificar patrones coherentes sobre los factores que más influyen en el error. Se ha verificado, por ejemplo, que las vistas laterales y traseras incrementan el error respecto a la frontal, y que tanto el tipo de acción como la articulación considerada son determinantes. Este análisis cualitativo ha validado las observaciones previas y ha añadido

transparencia al proceso, permitiendo una comprensión más profunda del comportamiento de los modelos evaluados.

En conjunto, los resultados permiten concluir que MHFormer constituye la alternativa más adecuada para tareas de estimación de pose 3D en entornos reales con múltiples perspectivas. MotionBERT puede resultar útil en escenarios donde se prioriza la cobertura de frames, aun a costa de una mayor tasa de error. Mediapipe, por su parte, presenta limitaciones importantes en contextos complejos, aunque su eficiencia lo convierte en una opción viable en aplicaciones con restricciones computacionales.

En conclusión, este trabajo ha contribuido tanto al desarrollo de un conjunto de datos original y de calidad como a la evaluación sistemática de modelos de estimación de pose 3D. Asimismo, se ha demostrado la utilidad de combinar métricas objetivas con herramientas de explicabilidad, lo cual ha permitido una interpretación más completa, robusta y transparente de los resultados obtenidos.

7.1. LIMITACIONES Y TRABAJO FUTURO

La principal limitación identificada en el trabajo se relaciona con la precisión en la localización absoluta del sujeto proporcionada por el traje Rokoko Smartsuit Pro II. Aunque el sistema incluye la unidad Coil Pro, diseñada para mejorar el geoposicionamiento mediante una unidad de referencia global, se ha observado un error sistemático en la posición absoluta del cuerpo. Este desajuste afecta especialmente a la proyección de los keypoints en los frames como se explica en el apartado 5.2.5. Este problema ha impedido realizar un ajuste espacial fiable entre ambas fuentes de datos. Si bien existe la posibilidad de aplicar un proceso de calibración posterior, este no ha podido llevarse a cabo por limitaciones temporales. Se prevé que futuras actualizaciones del firmware por parte del fabricante Rokoko aborden esta limitación.

Además, no se han incluido acciones que impliquen una separación simultánea de ambos pies del suelo (como saltos), debido a que el sistema requiere un reajuste posterior para mantener la coherencia de la posición corporal. Este ajuste tampoco ha podido realizarse por las mismas restricciones temporales, aunque también se espera que pueda solventarse mediante futuras mejoras del sistema.

Como líneas de trabajo futuro, se plantea abordar la corrección del error posicional mediante técnicas de ajuste posterior o refinamiento basado en visión por computador, lo que permitiría utilizar los datos del traje como supervisión efectiva para el entrenamiento de modelos de lifting. Adicionalmente, se propone como desafío investigar y desarrollar modelos de detección directa verdaderamente independientes de la cámara, capaces de inferir coordenadas 3D a partir de imágenes RGB sin

requerir parámetros externos. Este enfoque, aún en desarrollo, representa una vía prometedora para simplificar los flujos de trabajo y mejorar la robustez en entornos no controlados.

Otra dirección relevante consiste en aprovechar el sistema empleado para la generación de nuevos conjuntos de datos centrados en acciones específicas o situaciones de alta complejidad. La portabilidad y flexibilidad del traje Rokoko permiten registrar movimientos que suelen estar escasamente representados en datasets públicos, como actividades deportivas, procesos industriales, rutinas de rehabilitación o interacciones en entornos reales. Esta capacidad facilitaría la construcción de modelos especializados y contribuiría al enriquecimiento del ecosistema de datos en el ámbito de la estimación de la pose humana en 3D.

7.2. CONCLUSIONES

Este trabajo ha demostrado la viabilidad de construir y utilizar un sistema propio de captura de movimiento combinando un traje inercial con un sistema multicámara, como herramienta para el análisis de modelos de estimación de pose 3D. Se han evaluado tres modelos representativos del estado del arte (MediaPipe, MotionBERT y MHFormer), extrayendo conclusiones relevantes sobre su rendimiento ante distintas vistas, articulaciones, acciones y sujetos. El análisis se ha complementado con técnicas de explicabilidad que permiten interpretar los resultados de forma más precisa.

Se ha realizado un trabajo exhaustivo que ha involucrado múltiples aspectos técnicos, como la configuración del sistema de captura, la calibración del equipamiento, la sincronización temporal entre fuentes de datos y el tratamiento de información multimodal. Todo ello ha permitido un profundo entendimiento de diversos conceptos clave relacionados con la adquisición, procesamiento y análisis de datos para la estimación de pose humana.

Además de la construcción del conjunto de datos, se ha llevado a cabo su validación empírica mediante la aplicación de modelos de estimación de pose, identificando hallazgos que respaldan la hipótesis inicial sobre la necesidad de generar un dataset etiquetado con una variedad de actividades y capturado en un entorno no controlado. Esta aproximación aporta valor frente a los conjuntos tradicionales, generalmente limitados a entornos de laboratorio y condiciones idealizadas.

A pesar de ciertas limitaciones técnicas, especialmente relacionadas con la precisión del geoposicionamiento del traje y la falta de modelos abiertos adaptados a datos inerciales, el sistema desarrollado constituye una base sólida para futuras aplicaciones. Su capacidad para generar

datos específicos lo hace útil tanto para entrenar modelos especializados como para aplicaciones prácticas en ámbitos como la salud, el deporte o la interacción humano-máquina.

Este trabajo sienta las bases para una prometedora línea de investigación futura, con múltiples ramificaciones de gran interés, como la mejora de los algoritmos de fusión sensorial, la personalización de modelos de estimación de pose o el desarrollo de soluciones portables para el análisis de movimiento en tiempo real en entornos naturales.

8. OBJETIVOS DE DESARROLLO SOSTENIBLE

Los objetivos de este Trabajo Fin de Grado están alineados con los siguientes Objetivos de Desarrollo Sostenible (ODS) y metas, de la Agenda 2030:

- Objetivo 3: ODS 3: Salud y bienestar. La estimación de pose en 3D tiene aplicaciones directas en el ámbito de la salud y la biomecánica, facilitando el análisis del movimiento en estudios clínicos y deportivos.

- Objetivo 9: Industria, innovación e infraestructura. El trabajo propone una alternativa más accesible y económica a los métodos tradicionales de captura de movimiento basados en sistemas ópticos.

- Objetivo 12: Producción y consumo responsables. Al utilizar sensores inerciales en lugar de cámaras y sistemas de captura ópticos costosos, se reduce el consumo de materiales y energía en la implementación de laboratorios de captura de movimiento. Esto permite democratizar el acceso a la tecnología sin generar un impacto ambiental significativo.

- Objetivo 13: Acción por el clima: La reducción del uso de infraestructuras complejas y energéticamente demandantes en la investigación y desarrollo de modelos de estimación de pose contribuye a disminuir la huella de carbono.

9. BIBLIOGRAFÍA

Actor Profile. (s. f.). Rokoko. Recuperado 27 de abril de 2025, de https://support.rokoko.com/hc/en-us/articles/4410415403025-Actor-Profile

Award Winning Motion Capture Systems. (s. f.). Vicon. Recuperado 18 de marzo de 2025, de https://www.vicon.com/

Aznar-Gimeno, R., Perez-Lasierra, J. L., Pérez-Lázaro, P., Bosque-López, I., Azpíroz-Puente, M., Salvo-Ibáñez, P., Morita-Hernandez, M., Hernández-Ruiz, A. C., Gómez-Bernal, A., Rodrigalvarez-Chamarro, M. de la V., Alfaro-Santafé, J.-V., del Hoyo-Alonso, R., & Alfaro-Santafé, J. (2024). Gait-Based AI Models for Detecting Sarcopenia and Cognitive Decline Using Sensor Fusion. *Diagnostics*, *14*(24), 2886. https://doi.org/10.3390/diagnostics14242886

Bazarevsky, V., Grishchenko, I., Raveendran, K., Zhu, T., Zhang, F., & Grundmann, M. (2020). *BlazePose: On-device Real-time Body Pose tracking* (No. arXiv:2006.10204). arXiv. https://doi.org/10.48550/arXiv.2006.10204

Coil Pro FAQs. (s. f.). Rokoko. Recuperado 27 de abril de 2025, de https://support.rokoko.com/hc/en-us/articles/21598651434641-Coil-Pro-FAQs

Coil Pro—Known Issues. (s. f.). Rokoko. Recuperado 27 de abril de 2025, de https://support.rokoko.com/hc/en-us/articles/22166939492881-Coil-Pro-Known-Issues

GoPro HERO11 Black (cámara deportiva y subacuática). (s. f.). Recuperado 30 de mayo de 2025, de https://gopro.com/es/es/shop/cameras/hero11-black/CHDHX-111-master.html

Intuitive and affordable motion capture tools for character animation. (s. f.). Recuperado 27 de abril de 2025, de https://support.rokoko.com/hc/en-us/articles/17672094977809-Smartsuit-Pro-II-Setup-Guide

Ionescu, C., Papava, D., Olaru, V., & Sminchisescu, C. (2014). Human3.6M: Large Scale Datasets and Predictive Methods for 3D Human Sensing in Natural Environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *36*(7), 1325-1339. https://doi.org/10.1109/TPAMI.2013.248

Joo, H., Simon, T., Li, X., Liu, H., Tan, L., Gui, L., Banerjee, S., Godisart, T., Nabbe, B., Matthews, I., Kanade, T., Nobuhara, S., & Sheikh, Y. (2019). Panoptic Studio: A Massively Multiview System for Social Interaction Capture. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *41*(1), 190-204. https://doi.org/10.1109/TPAMI.2017.2782743

Li, W. (2025). *Vegetebird/MHFormer* [Python]. https://github.com/Vegetebird/MHFormer (Obra original publicada en 2021)

Lin, T.-Y., Maire, M., Belongie, S., Bourdev, L., Girshick, R., Hays, J., Perona, P., Ramanan, D., Zitnick, C. L., & Dollár, P. (2015). *Microsoft COCO: Common Objects in Context* (No. arXiv:1405.0312). arXiv. https://doi.org/10.48550/arXiv.1405.0312

Lugaresi, C., Tang, J., Nash, H., McClanahan, C., Uboweja, E., Hays, M., Zhang, F., Chang, C.-L., Yong, M. G., Lee, J., Chang, W.-T., Hua, W., Georg, M., & Grundmann, M. (2019). *MediaPipe: A Framework for Building Perception Pipelines* (No. arXiv:1906.08172). arXiv. https://doi.org/10.48550/arXiv.1906.08172

Mehta, D., Rhodin, H., Casas, D., Fua, P., Sotnychenko, O., Xu, W., & Theobalt, C. (2017). *Monocular 3D Human Pose Estimation In The Wild Using Improved CNN Supervision* (No. arXiv:1611.09813). arXiv. https://doi.org/10.48550/arXiv.1611.09813

Motion Capture Systems. (s. f.). OptiTrack. Recuperado 18 de marzo de 2025, de http://www.optitrack.com/index.html

Neupane, R. B., Li, K., & Boka, T. F. (2024). A survey on deep 3D human pose estimation. *Artificial Intelligence Review*, *58*(1), 24. https://doi.org/10.1007/s10462-024-11019-3

OpenCV: Camera Calibration and 3D Reconstruction. (s. f.). Recuperado 29 de mayo de 2025, de https://docs.opencv.org/4.x/d9/d0c/group__calib3d.html

Pavllo, D., Feichtenhofer, C., Grangier, D., & Auli, M. (2019). *3D human pose estimation in video with temporal convolutions and semi-supervised training* (No. arXiv:1811.11742). arXiv. https://doi.org/10.48550/arXiv.1811.11742

Perception Neuron Series | Noitom Motion Capture Systems. (s. f.). Recuperado 18 de marzo de 2025, de https://www.noitom.com/perception-neuron-series

Procrustes analysis. (2025). En *Wikipedia*. https://en.wikipedia.org/w/index.php?title=Procrustes_analysis&oldid=1289748187

Sigal, L., Balan, A. O., & Black, M. J. (2010). HumanEva: Synchronized Video and Motion Capture Dataset and Baseline Algorithm for Evaluation of Articulated Human Motion. *International Journal of Computer Vision*, *87*(1), 4-27. https://doi.org/10.1007/s11263-009-0273-6

Smartsuit Pro II - Quality body motion capture in one simple mobile mocap suit. (s. f.). Recuperado 18 de marzo de 2025, de https://www.rokoko.com/products/smartsuit-pro

Tekulve, W. (2025). *Tekulvw/bvh-converter* [Python]. https://github.com/tekulvw/bvh-converter (Obra original publicada en 2016)

Trumble, M., Gilbert, A., Hilton, A., & Collomosse, J. (2018). *Deep Autoencoder for Combined Human Pose Estimation and body Model Upscaling* (No. arXiv:1807.01511). arXiv. https://doi.org/10.48550/arXiv.1807.01511

Von Marcard, T., Henschel, R., Black, M. J., Rosenhahn, B., & Pons-Moll, G. (2018). Recovering Accurate 3D Human Pose in the Wild Using IMUs and a Moving Camera. En V. Ferrari, M. Hebert, C. Sminchisescu, & Y. Weiss (Eds.), *Computer Vision – ECCV 2018* (Vol. 11214, pp. 614-631). Springer International Publishing. https://doi.org/10.1007/978-3-030-01249-6_37

Xsens MVN Animate—Motion Capture Software for Professionals. (s. f.). Recuperado 18 de marzo de 2025, de https://www.movella.com/products/motion-capture/xsens-mvn-animate

Xu, Y., Wang, W., Liu, T., Liu, X., Xie, J., & Zhu, S.-C. (2022). Monocular 3D Pose Estimation via Pose Grammar and Data Augmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *44*(10), 6327-6344. https://doi.org/10.1109/TPAMI.2021.3087695

Ye, H., Zhu, W., Wang, C., Wu, R., & Wang, Y. (2022). *Faster VoxelPose: Real-time 3D Human Pose Estimation by Orthographic Projection* (No. arXiv:2207.10955). arXiv. https://doi.org/10.48550/arXiv.2207.10955

Zhu, W., Ma, X., Liu, Z., Liu, L., Wu, W., & Wang, Y. (2023). *MotionBERT: A Unified Perspective on Learning Human Motion Representations* (No. arXiv:2210.06551). arXiv. https://doi.org/10.48550/arXiv.2210.06551

Relación de documentos

(X) Memoria 87 páginas
(_) Anexos 8 páginas

La Almunia, a 01 de 07 de 2025

Firmado: David Polo Llimós