CEA Comité Español de Automática
XLV JORNADAS AUTOMÁTICA MÁLAGA 2024
UNIVERSIDADE DA CORUÑA

Jornadas de Automática

Sistema multi-cámara de estimación de pose sin marcadores para pHRI

Asensio-Huonder, S.[a,*], Fernández-Carmona, M.[b], Arévalo-Espejo, V.[a], Urdiales, A.C.[b], Gómez-de-Gabriel, J.M.[a]

[a]*Dpto. de Ingeniería de Sistemas y Automática, Universidad de Málaga, 29071 Málaga, España.*
[b]*Dpto. de Tecnología Electrónica, Universidad de Málaga, 29071 Málaga, España.*

To cite this article: Asensio-Huonder, S., Fernández-Carmona, M., Arévalo-Espejo, V., Urdiales, A.C., Gómez-de-Gabriel, J.M. 2024. Markerless multi-camera pose estimation system for pHRI. Jornadas de Automática, 45. https://doi.org/10.17979/ja-cea.2024.45.10909

Resumen

Este trabajo presenta un sistema basado en visión que utiliza redes neuronales para la estimación de poses humanas en 3D. La solución desarrollada identifica en el sujeto analizado 18 "puntos clave" o *keypoints* mediante cuatro cámaras RGB calibradas. La utilización de múltiples cámaras permite superar problemas inherentes al uso de una sola cámara RGBD/estéreo como la pérdida de *keypoints* por la existencia de oclusiones o una mayor incertidumbre en la estimación de la profundidad, proporcionando una base robusta para futuras investigaciones y aplicaciones en campos como la rehabilitación física. Asimismo, se ha registrado y puesto a disposición de la comunidad la posición 3D de los *keypoints* identificados durante la realización de seis ejercicios distintos enfocados en el movimiento del brazo derecho del sujeto. Este trabajo contribuye a la literatura actual ofreciendo un sistema novedoso para la obtención de la pose humana y demuestra la viabilidad de nuestra metodología, abriendo nuevas vías para investigaciones futuras en el contexto del pHRI.

Palabras clave: Programación y Visión, Diseño, modelado y análisis de HMS, Tecnología asistiva e ingeniería de rehabilitación.

Markerless multi-camera pose estimation system for pHRI

Abstract

This paper presents a vision-based system that utilizes neural networks for estimating human poses in 3D. The developed solution identifies 18 keypoints on the subject using four calibrated RGB cameras. The use of multiple cameras overcomes problems inherent to the use of a single RGBD/stereo camera, such as the loss of keypoints due to occlusions or increased uncertainty in depth estimation, providing a robust foundation for future research and applications in fields such as physical rehabilitation. Furthermore, the 3D position of the identified keypoints was recorded and made available to the community during the performance of six different exercises focused on the movement of the subject's right arm. This work contributes to the current literature by offering a novel system for obtaining human pose and demonstrates the feasibility of our methodology, paving the way for future research in the context of pHRI.

Keywords: Programming and Vision, Design, modeling and analysis of HMS, Assistive technology and rehabilitation engineering.

1. Introducción

La estimación de la pose humana (HPE) es una tarea que emplea técnicas de visión por computadora para determinar la configuración del cuerpo humano en una imagen dada o en una secuencia de imágenes. Esta es una tarea importante en el campo de la visión por computador y se utiliza en una amplia gama de dominios científicos y de consumo. Algunos ejemplos incluyen: interacción humano-computadora (HCI), el movimiento humano puede proporcionar interfaces

*Autor para correspondencia: asensioh.santiago@uma.es

naturales con las que los ordenadores pueden ser controlados mediante gestos humanos o pueden reconocer lenguajes de signos (Moryossef et al., 2021); interacción humano-robot (HRI): en entornos domésticos, y especialmente en situaciones de asistencia, es esencial que un robot pueda percibir la pose del cuerpo humano para interactuar de manera más efectiva (Droeschel and Behnke, 2011). En el ámbito del análisis de rendimiento deportivo, los movimientos de los atletas se estudian en profundidad desde múltiples vistas y, como resultado, los sistemas multi-cámara de estimación de poses pueden ayudar a analizar estas acciones (Unzueta et al., 2014; Desmarais et al., 2021).

Hay pocos recursos disponibles en lo que se refiere a la estimación de la pose humana tridimensional; se pueden encontrar algunos trabajos que emplean exoesqueletos para la estimación de pose humana (Gilbert et al., 2019), así como trabajos en los que se colocan marcadores sobre el sujeto y la pose se recupera mediante la localización de dichos marcadores mediante un sistema de visión (Lo Presti and La Cascia, 2016). Estos sistemas proporcionan estimaciones de pose con gran precisión, pero poseen varios inconvenientes, tales como la necesidad de disponer de elementos adicionales al sistema de visión para la recogida de datos, resultando intrusivos en la medida en que se adhieren elementos al sujeto. Gracias al reciente desarrollo de las redes neuronales convolucionales (CNNs), la investigación en la estimación de poses humanas empleando métodos *markerless* que no requieren elementos emplazados sobre el sujeto ha experimentado avances significativos recientemente (Wei et al., 2016; Newell et al., 2016; Xiao et al., 2018a; Slembrouck et al., 2020). Por ejemplo, la red de alta resolución *HRNet* (Sun et al., 2019) utiliza múltiples ramas de resolución a lo largo de toda la red y ha logrado un rendimiento líder en conjuntos de datos públicos.

Las técnicas descritas anteriormente extraen "puntos clave" o *keypoints* (tales como muñeca, rodilla, etc.) a partir de imágenes RGB por lo que poseen el inconveniente de no ser robustas ante la presencia de oclusiones. Luego, a pesar de estos avances tecnológicos y del lanzamiento de nuevos conjuntos de datos que han resultado en un creciente interés de la comunidad científica en este campo, la estimación de la pose humana sigue siendo un problema abierto con varios desafíos.

En este trabajo se presenta un sistema multi-cámara para la estimación de poses humanas en 3D mediante un procedimiento *markerless* no intrusivo. Para identificar los *keypoints* se recurre a redes neuronales entrenadas con el *dataset* público COCO18 (Lin et al., 2014). El sistema fusiona la información proveniente de cuatro cámaras para reducir la incidencia de las oclusiones y proporcionar una mejor estimación de la pose 3D del sujeto. Adicionalmente, se incluye y describe un *dataset*[1] preliminar para la validación del sistema, con los *keypoints* 3D extraídos durante la realización de seis ejercicios diferentes realizados con el brazo derecho del sujeto. Los datos han sido almacenados en ficheros compatibles con ROS (*Robotic Operating System*) que han sido puestos a disposición de la comunidad. Los detalles sobre el montaje del sistema[2] también han sido puestos a disposición de la comunidad.

2. Descripción del sistema propuesto

En esta sección se describe el sistema multi-cámara propuesto para la estimación de la pose 3D humana. En primer lugar se detalla el equipamiento informático y sensorial utilizado en la toma de datos, su disposición en el espacio de trabajo y el *software* utilizado para calibrarlos extrínsecamente. A continuación, se describe el *software* encargado de extraer los *keypoints* del cuerpo del sujeto a partir de imágenes RGB. Para terminar se describe el procedimiento desarrollado para determinar la pose 3D de los *keypoints* identificados en el paso anterior.

2.1. Equipamiento utilizado

En la toma de datos se han utilizado cuatro cámaras *RGBD Intel RealSense*, dos D435 y dos D415, unidas a una estructura rectangular de 2.0×1.5 m colocada verticalmente sobre el suelo (ver figura 1.b). Para fijarlas rígidamente a la estructura y orientarlas hacia el espacio de trabajo se han diseñado unos soportes *ad-hoc* impresos en 3D (ver detalle en la misma figura). Esta disposición busca, primero, facilitar la calibración extrínseca de los sensores, asegurando cierto solape entre imágenes, y segundo, contar siempre con suficiente información visual para extraer *keypoints* fiables en caso de que alguna parte del sujeto quede ocluida o se produzca un fallo en la detección en alguna de las cámaras.

La ejecución del *software* de detección de *keypoints* se realiza de forma distribuida en dos NVIDIA *Jetson Xavier*. Cada una de ellas procesa la información RGB proveniente de dos cámaras y transfiere la información relativa a los *keypoints* detectados a un potente PC, que será el encargado de ejecutar el *software* destinado a determinar sus poses 3D con respecto a la base del brazo *Franka Research 3* colocado delante de la estructura rectangular (ver figura 1.b). Los tres dispositivos se conectan a un conmutador *ethernet* de alta velocidad y sus relojes internos han sido previamente sincronizados utilizando un servidor NTP (*Network Time Protocol*) local que corre en el PC. El objetivo es reducir la latencia de las comunicaciones y garantizar la integridad temporal de los datos transferidos.

2.2. Calibración extrínseca de los sensores

Para poder determinar de forma precisa la pose 3D de los *keypoints* extraídos con respecto al sistema de referencia del mundo es necesario realizar una calibración intrínseca y extrínseca de los sensores utilizados. La intrínseca permite determinar la matriz de proyección de cada sensor (incluidos los parámetros de distorsión) y la extrínseca consiste en determinar su posición con respecto al sistema de referencia del mundo. Los sensores utilizados en este trabajo vienen calibrados de fábrica, por lo que solo se ha realizado la segunda.

El *software* utilizado para calibrar los cuatro sensores, aunque ligeramente modificado para adaptarse a nuestra configuración, forma parte del proyecto *Open PTrack* (Munaro et al., 2016), una solución *open-source* para el seguimiento de personas en interiores a partir de redes de sensores RGB. El citado *software* utiliza un patrón *checkboard* similar al utilizado típicamente en la calibración intrínseca, pero de mayores

[1] https://www.kaggle.com/datasets/jesusgomezdegabriel/human-upper-limb-joints-with-vision-only-for-phri
[2] https://github.com/TaISLab/pHRI-Vision-Robot-Workstation

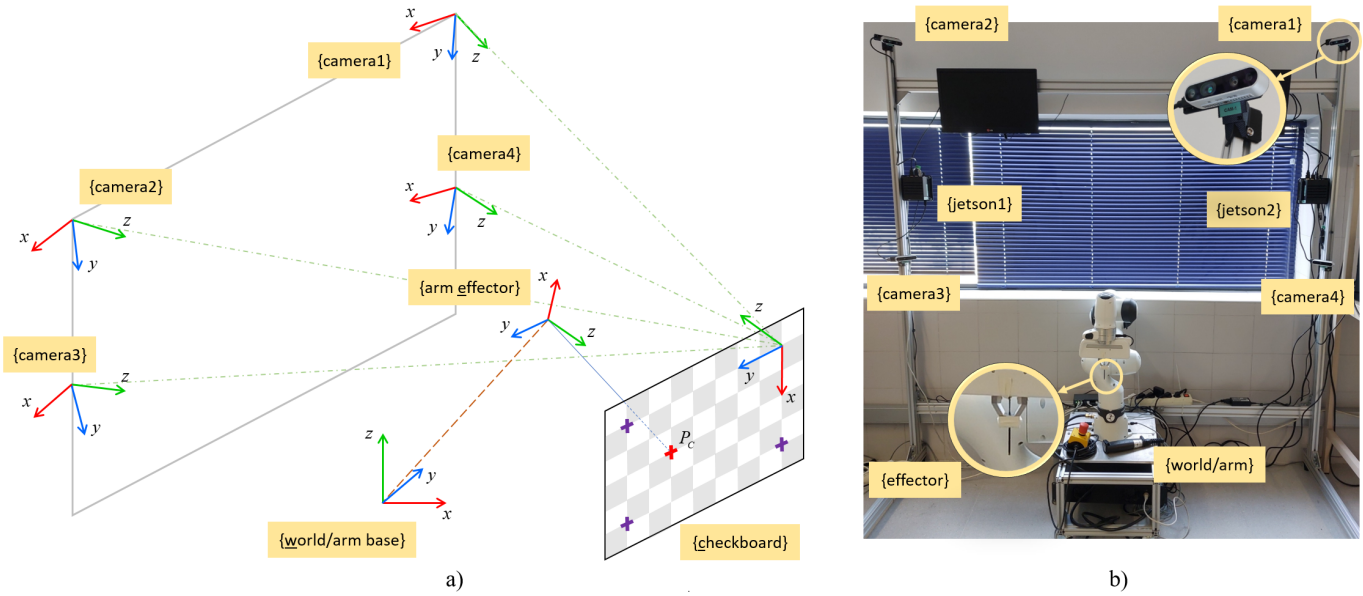

Figura 1: a) Sistemas de referencia de los distintos elementos que integran el sistema de captura: sensores *Intel RealSense*, el brazo *Franka Research 3* (*world*) y el patrón de calibración. b) Fotografía de la estructura utilizada en la toma de datos. Los sensores {camera1} y {camera2} se corresponden al modelo D435 y {camera3} y {camera4} al modelo D415.

dimensiones. Para realizar la calibración se ha utilizado un patrón de 10×8 celdas con un tamaño de celda de 8×8 cm colocado delante de la estructura rectangular de forma que sea observado por las cuatro cámaras. Durante el proceso de toma de datos no se puede mover el patrón.

Una vez finalizado el proceso de calibración, el *software* proporciona dos conjuntos de datos: la calibración de tres cámaras con respecto a una de ellas (se puede elegir qué cámara antes de iniciar el procedimiento) y la calibración de las cuatro cámaras con respecto al sistema de referencia del patrón utilizado, cuyo origen está situado en el vértice interior de la celda blanca situada en la esquina superior derecha del tablero (ver figura 1.a). Puesto que el objetivo es referenciar todas las cámaras con respecto a la base del brazo, se ha optado por la segunda. El procedimiento seguido para establecer esta última relación geométrica consta de los siguientes pasos:

a) Se eligen al menos tres puntos no colineales distribuidos sobre el mismo patrón utilizado en el paso anterior, que permanece estático delante de las cuatro cámaras. Bastaría con seleccionar el origen del sistema de referencia del patrón y los vértices internos de las otras celdas situadas en las esquinas. Si bien, se ha optado por seleccionar un número mayor con el fin de lograr un ajuste más preciso.
b) Seguidamente, se opera el brazo manualmente hasta que el extremo del efector final toque cada uno de los puntos elegidos, obteniendo de este modo sus posiciones 3D con respecto a la base. Es importante que el brazo no mueva el patrón durante el proceso de captura de posiciones. Si esto ocurriera, habría que repetir el proceso de calibración desde el inicio.
c) Finalmente, se determina la transformación geométrica (rotación y traslación) que relaciona ambos sistemas mediante un procedimiento de optimización que minimice los errores de ajuste de los puntos 3D seleccionados.

La figura 1 muestra esquemáticamente la ubicación de los cuatro sensores *Intel Realsense*, junto con sus sistemas de referencia; la posición del sistema de referencia del mundo, situado en la base del brazo manipulador; y el tablero utilizado en el proceso de calibración, junto a sus sistemas de referencia. Recuérdese que, aunque el manipulador no interviene en la grabación de datos, sí que está previsto que se utilice en futuros trabajos, de ahí que el sistema de referencia del mundo esté situado en su base.

2.3. *Extracción de keypoints*

La extracción robusta y en tiempo real de *keypoints* se realiza mediante el proyecto *open-source* de NVIDIA *TRT pose* (Yato and Welsh, 2021), el cual emplea métodos de detección de postura en imágenes RGB utilizando *deep learning* (Cao et al., 2017; Xiao et al., 2018b). Su principio de funcionamiento es similar al de librerías como *OpenPose* (Cao et al., 2021), pero su optimización para plataformas de NVIDIA (como las *Jetson Xavier* utilizadas aquí) le proporciona un mayor rendimiento, ofreciendo tasas de hasta 251 *frames* por segundo. *TRT pose* utiliza *TensorRT* para optimizar y mejorar el rendimiento de los métodos de *deep learning* en GPUs de NVIDIA y proporciona dos modelos de redes neuronales convolucionales (CNNs) entrenadas en el *dataset* COCO18 (Lin et al., 2014), un conjunto de imágenes RGB etiquetadas con las posiciones de 18 "puntos de interés" humanos.

Para la inferencia en tiempo real de los *keypoints* se parte del modelo optimizado para *tensorRT*. Las imágenes son reescaladas al tamaño utilizado por el modelo (254×254), normalizadas y convertidas a tensores para el proceso de inferencia. El modelo proporciona como salida un mapa de probabilidad de los *keypoints*, así como un campo de afinidad (utilizado para expresar las relaciones entre *keypoints*) referido a la imagen de entrada. Postprocesando estas dos componentes, es posible establecer una lista ordenada de *keypoints* detectados en la

imagen conforme a la numeración establecida en la figura 3. Las posiciones de dicho conjunto de puntos representan rayos que apuntan a cada uno de los *keypoints* y están naturalmente referidas a las coordenadas de la cámara (ver figura 2).

2.4. *Estimación de la posición 3D*

La etapa final del sistema de estimación de la pose humana propuesto pasa por determinar la posición 3D de los *keypoints* detectados. En este trabajo se estima por triangulación a partir de los rayos generados (uno por cada cámara que observa el *keypoint*) en la etapa previa, descartando la información de profundidad proporcionada por los propios sensores RGBD por ser muy ruidosa. Este enfoque requiere que cada *keypoint* sea observado por, al menos, dos cámaras; si no lo es, se le asigna la posición $(0, 0, 0)$. El hecho de que se disponga de cuatro rayos para estimar la posición 3D permite lidiar con situaciones en las que partes del cuerpo del sujeto quedan circunstancialmente ocluidas desde un punto de vista, pero no desde los otros tres.

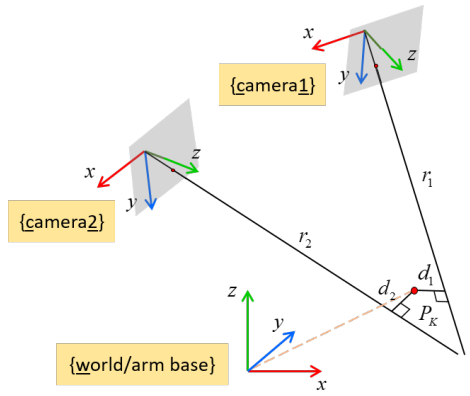

Figura 2: Representación esquemática del proceso de estimación de la pose 3D de una marca a partir de los rayos proporcionados por el sistema de extracción de *keypoints*.

En condiciones ideales, la posición 3D de cualquier *keypoint* vendría dada por la intersección de todos sus rayos. Sin embargo, la información sensorial está afectada de ruido, siendo perfectamente posible que ninguno de los cuatro rayos intersequen. En este trabajo se propone ubicarlo en aquella posición del espacio que minimice la distancia normal a cada uno ellos (ver figura 2). Por tanto, atendiendo a lo dicho anteriormente, la estimación de la posición 3D del *keypoint* i-ésimo, P_i, se podría definir formalmente como el siguiente problema de optimización:

$$\hat{P}_i = \arg\min_{P_i} \sum_{j=1}^{n} \frac{\left\| \overrightarrow{(P_i\, O_j)} \times \vec{r}_{i_j} \right\|}{\left\| \vec{r}_{i_j} \right\|} \qquad (1)$$

donde O_j es el origen del sistema de referencia de la cámara j-ésima; $\vec{r}_{i_j}$ el vector director del rayo j-ésimo que contiene el *keypoint* i-ésimo; y n el número de cámaras que lo han observado. En la formulación propuesta se asume que tanto las coordenadas como los vectores directores de los rayos están expresados con respecto al sistema de referencia del mundo.

El problema de minimización planteado en (1) se resuelve mediante el algoritmo de optimización no lineal *Trust Region Reflective* (o TRF) (Branch et al., 1999), y el valor que se le asigna inicialmente a P_i es el promedio de los orígenes de los sistemas de referencia de cada una de las cámaras desde las que ha sido observado el *keypoint*. Este proceso se realiza para todos y cada uno de los *keypoints* detectados y es fácilmente paralelizable.

3. Estructura de los datos generados por el sistema

En esta sección se describe brevemente la estructura de los datos generados por el sistema de estimación de pose humana propuesto. El sistema visualiza y registra las poses 3D del sujeto que se encuentra en el espacio de trabajo, manteniendo la numeración de los *keypoints* establecida en la figura 3. Las coordenadas, como se indica en la sección 2.4, están expresadas con respecto a la base del manipulador en pos de facilitar el uso conjunto del manipulador y el sistema multi-cámara propuesto en futuras aplicaciones de rehabilitación.

El lector puede dirigirse al repositorio proporcionado en la sección 1 para obtener más detalles sobre la estructura de los datos guardados y cómo utilizarlos. En el citado repositorio también está disponible un *dataset* preliminar con las posiciones 3D estimadas de los *keypoints* de un sujeto que lleva a cabo una serie de movimientos con el brazo derecho. Nótese que, aunque en la sección 4 se ilustre la viabilidad del sistema con los 3 *keypoints* asociados al brazo, en aplicaciones de rehabilitación asistida por manipuladores, es preciso conocer en todo momento la pose del cuerpo del sujeto con objeto de evitarle daños, de ahí que se registren los 18 *keypoints*. Los ficheros de datos generados están en formato ROS. Estos ficheros, denominados *rosbags*, son empleados comúnmente por la comunidad para almacenar mensajes ROS en formato binario, resultando útiles en la grabación de datos provenientes de sensores, actuadores y algoritmos durante el funcionamiento del sistema y permitiendo su posterior procesamiento y análisis.

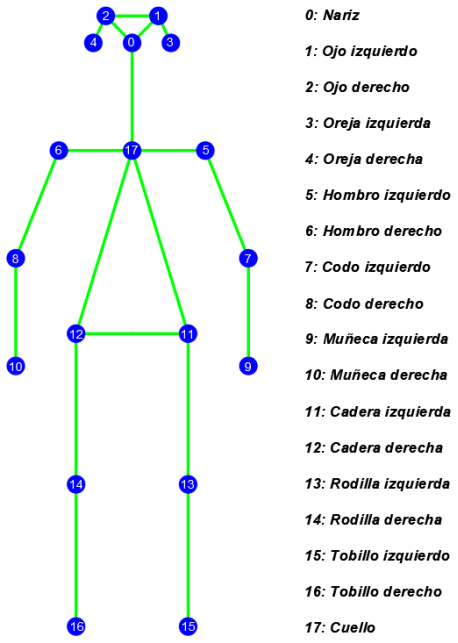

Figura 3: Localización y etiquetado de los 18 *keypoints* identificados por el sistema de estimación de la pose humana propuesto (formato COCO18).

4. Experimentos y resultados

En esta sección se describen los experimentos realizados y el análisis de los resultados en cuanto a la varianza en las distancias entre articulaciones adyacentes y representación de las trayectorias efectuadas por las posiciones de la muñeca. Los resultados han sido obtenidos a partir de la pose de los *keypoints* correspondientes a la muñeca, el codo y el hombro del brazo derecho del sujeto, haciendo uso de los datos grabados.

4.1. Experimentos

Los seis tipos de movimientos realizados por uno de los voluntarios se recogen a continuación. Todos ellos se realizan evitando que el brazo manipulador ocluya parte de la escena, y manteniéndose dentro del espacio de trabajo tanto del sistema de visión como del manipulador. La selección de los ejercicios se ha elegido tomando como referencia trabajos existentes en el ámbito de la rehabilitación mediante robots (Macovei and Doroftei, 2016; Qie et al., 2022). La ejecución se realiza sin la ayuda de marcas visuales, ni objetivos de tiempo o dimensiones de las trayectorias tal como las realizaría un humano de forma natural, como se observa en la figura 4.

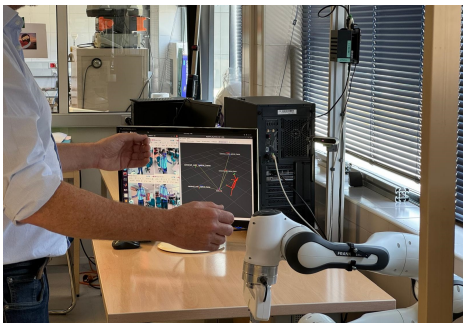

Figura 4: Uso del sistema de visión multi-cámara para captura de movimientos no intrusiva del humano en interacción con el espacio de trabajo del robot.

Ejercicio 1: Movimiento del brazo derecho siguiendo una trayectoria circular. El ejercicio consiste en trazar dos veces un círculo en sentido contrario a las agujas del reloj, de forma paralela al suelo.

Ejercicio 2: Movimiento del brazo derecho siguiendo una trayectoria cuadrada. En este ejercicio el voluntario traza una trayectoria cuadrada en sentido contrario a las agujas del reloj y de forma paralela al suelo, completando dos vueltas.

Ejercicio 3: Flexión vertical del codo. Se flexiona el codo derecho 2 veces manteniendo el brazo en posición vertical; el movimiento inicia con el brazo extendido, el cual es flexionado hasta formar 90 grados con el antebrazo.

Ejercicio 4: Flexión horizontal del codo. Se flexiona el codo derecho 3 veces manteniendo el brazo en posición horizontal.

Ejercicio 5: Movimiento de subida y bajada del brazo derecho extendido. El sujeto inicia el movimiento con el brazo próximo a la cadera, completamente extendido, y levanta el brazo derecho hasta situarlo de forma paralela al suelo, llevando a cabo este movimiento 2 veces.

Ejercicio 6: Aplaudir. La persona entrechoca las manos 4 veces.

Estos experimentos están aprobados por el Comité de Ética de la UMA con número de protocolo CEUMA 7-2023-H.

4.2. Resultados

En esta sección se presentan resultados haciendo uso de los *keypoints* 3D correspondientes al hombro (k_6), codo (k_8) y muñeca (k_{10}) del brazo derecho del sujeto. En primer lugar, se muestra la varianza en la distancia hombro-codo y codo-muñeca mediante gráficos *box plot*. A continuación, se muestran las trayectorias efectuadas por la muñeca del voluntario para los ejercicios 1 y 2 (ver sección 4.1). Finalmente, se representa la evolución del valor angular de la articulación del codo derecho durante cada ejercicio.

Las figuras 5 y 6 ilustran la precisión en la detección de los *keypoints* del brazo derecho del sujeto por medio del cómputo de la distancia entre articulaciones consecutivas en cada ejercicio. La varianza posee una sigma inferior a 4 cm en ambos casos y con valores medios de distancia entre articulaciones similares. Las trayectorias trazadas por la muñeca derecha del sujeto durante los ejercicios 1 (trayectoria circular) y 2 (trayectoria cuadrada) son semejantes a las formas geométricas esperadas (ver figura 7).

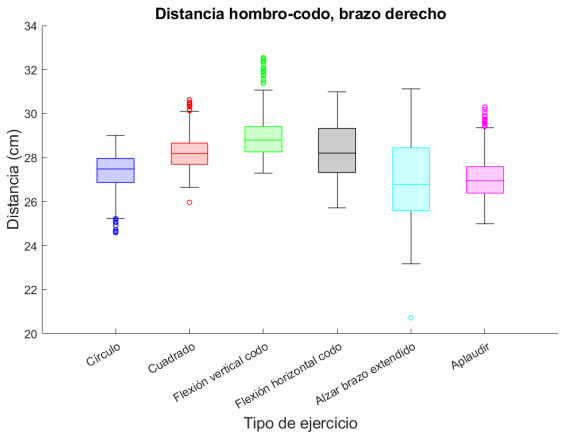

Figura 5: Distancia entre las articulaciones del hombro y el codo del brazo derecho.

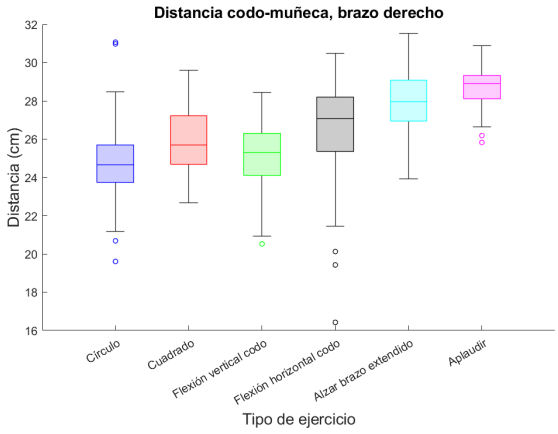

Figura 6: Distancia entre las articulaciones del codo y de la muñeca del brazo derecho.

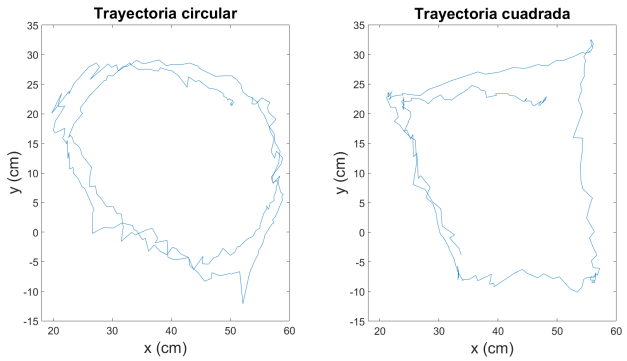

Figura 7: Trayectoria de la muñeca derecha del sujeto durante los ejercicios 1 y 2 (ver la sección 4.1).

En la figura 8 se muestran los valores angulares calculados para la articulación del codo derecho del sujeto. La consistencia de los valores angulares del codo derecho con los movimientos ejecutados por el sujeto refleja la capacidad del sistema para la identificación de poses humanas.

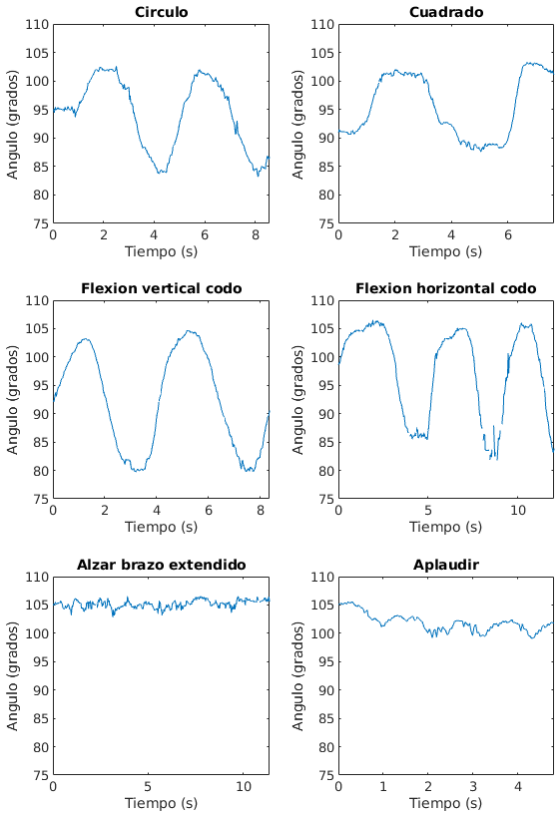

Figura 8: Valor angular de la articulación del codo derecho del voluntario.

5. Conclusiones

El presente trabajo introduce un sistema multi-cámara sin marcadores para la estimación de poses humanas en 3D, diseñado para aplicaciones de interacción humano-robot en ambientes de rehabilitación. Los resultados muestran la efectividad del sistema y del método de calibración extrínseca empleado. Está previsto trabajar en la generación de un *groud-truth*, así como la extensión del conjunto de datos grabados con nuevos ejercicios que involucren el movimiento de varias articulaciones del sujeto y en los que se haga uso de los 18 *keypoints* detectados por el sistema.

Agradecimientos

Este trabajo ha sido financiado por el proyecto de Generación de Conocimiento PID2021-127221OB-I00, del Ministerio de Ciencia, Innovación y Universidades.

Referencias

Branch, M., Coleman, T., li, Y., 1999. A subspace, interior, and conjugate gradient method for large-scale bound-constrained minimization problems. SIAM Journal on Scientific Computing 21 (1), 1–23.

Cao, Z., Hidalgo, G., Simon, T., Wei, S., Sheikh, Y., jan 2021. OpenPose: Realtime multi-person 2D pose estimation using part affinity fields. IEEE Transactions on Pattern Analysis and Machine Intelligence 43 (1), 172–186.

Cao, Z., Simon, T., Wei, S.-E., Sheikh, Y., 2017. Realtime multi-person 2D pose estimation using part affinity fields. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1302–1310.

Desmarais, Y., Mottet, D., Slangen, P., Montesinos, P., 2021. A review of 3D human pose estimation algorithms for markerless motion capture. Computer Vision and Image Understanding 212, 103275.

Droeschel, D., Behnke, S., 2011. 3D body pose estimation using an adaptive person model for articulated ICP. In: Intelligent Robotics and Applications. Springer Berlin Heidelberg, Berlin, pp. 157–167.

Gilbert, A., Trumble, M., Malleson, C., Hilton, A., Collomosse, J., 2019. Fusing visual and inertial sensors with semantics for 3D human pose estimation. International Journal of Computer Vision 127, 381–397.

Lin, T., Maire, M., Belongie, S. J., Bourdev, L. D., Girshick, R. B., Hays, J., Perona, P., Ramanan, D., Doll'a r, P., Zitnick, C. L., June 2014. Microsoft COCO: common objects in context. In: 13th European Conference on Computer Vision (ECCV). Zurich, pp. 740–755.

Lo Presti, L., La Cascia, M., 2016. 3d skeleton-based human action classification: A survey. Pattern Recognition 53, 130–147.

Macovei, S., Doroftei, I., aug 2016. A short overview of upper limb rehabilitation devices. IOP Conference Series: Materials Science and Engineering 145 (5), 052014.

Moryossef, A., Tsochantaridis, I., Dinn, J., Camgoz, N. C., Bowden, R., Jiang, T., Rios, A., Muller, M., Ebling, S., June 2021. Evaluating the immediate applicability of pose estimation for sign language recognition. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops. pp. 3434–3440.

Munaro, M., Basso, F., Menegatti, E., 2016. OpenPTrack: Open source multi-camera calibration and people tracking for RGB-D camera networks. Robotics and Autonomous Systems 75, 525–538.

Newell, A., Yang, K., Deng, J., Sept 2016. Stacked hourglass networks for human pose estimation. In: European Conference on Computer Vision (ECCV). Cham, pp. 483–499.

Qie, X., Kang, C., Zong, G., Chen, S., 2022. Trajectory planning and simulation study of redundant robotic arm for upper limb rehabilitation based on back propagation neural network and genetic algorithm. Sensors 22 (11).

Slembrouck, M., Luong, H., Gerlo, J., Schütte, K., Van Cauwelaert, D., De Clercq, D., Vanwanseele, B., Veelaert, P., Philips, W., 2020. Multiview 3D markerless human pose estimation from OpenPose skeletons. In: Advanced Concepts for Intelligent Vision Systems. Cham, pp. 166–178.

Sun, K., Xiao, B., Liu, D., Wang, J., June 2019. Deep high-resolution representation learning for human pose estimation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, pp. 5686–5696.

Unzueta, L., Goenetxea, J., Rodriguez, M., Linaza, M. T., Sept 2014. Viewpoint-dependent 3D human body posing for sports legacy recovery from images and video. In: 22nd European Signal Processing Conference. Lisbon, p. 361–365.

Wei, S.-E., Ramakrishna, V., Kanade, T., Sheikh, Y., June 2016. Convolutional pose machines. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, pp. 4724–4732.

Xiao, B., Wu, H., Wei, Y., Sept 2018a. Simple baselines for human pose estimation and tracking. In: European conference on computer vision (ECCV). Munich, pp. 466–481.

Xiao, B., Wu, H., Wei, Y., Sept 2018b. Simple baselines for human pose estimation and tracking. In: European Conference on Computer Vision (ECCV). Munich, pp. 472–487.

Yato, C., Welsh, J., 2021. trt_pose. https://github.com/NVIDIA-AI-IOT/trt_pose.