Estimación de pose y visión artificial: medir el movimiento con criterios biomecánicos

Autoría: EDUFISICA
Publicación: 2026-10-11
Revisión editorial: 2026-10-15
Lectura estimada: 28 minutos

https://edufisica.com/biomecanica/vision-artificial-deep-learning-pose-estimation/

Resumen

La estimación de pose añade puntos y segmentos a una grabación, pero interpretar ese dibujo exige conocer cómo se obtuvo. Esta lectura diferencia detección, estimación y seguimiento, y relaciona cada operación con las preguntas del análisis biomecánico. A partir de ejemplos matemáticos, explica la proyección de la cámara, la calibración, la reconstrucción tridimensional, los ángulos y el efecto del ruido al calcular velocidades. Después examina la validación de modelos, las oclusiones y la diferencia entre describir un movimiento y estimar las fuerzas que intervienen. El recorrido termina con decisiones para preparar una captura docente, comparar registros, reconocer límites y tratar con cuidado las grabaciones escolares. Las investigaciones citadas se distinguen de los ejemplos didácticos propios.

Palabras clave: estimación de pose; visión artificial; análisis del movimiento; calibración; cinemática; error de medición

Qué es la estimación de pose y qué aporta a la biomecánica

Una grabación muestra colores y formas en una sucesión de imágenes. La estimación de pose añade una representación de puntos corporales, como hombros, caderas o rodillas, a partir de esos datos. Su utilidad biomecánica depende de qué puntos estima, cómo los localiza y qué pregunta pretende responder el análisis.

Conviene separar tres expresiones que suelen utilizarse como equivalentes. La **detección de personas** localiza una región de la imagen donde aparece alguien. La **estimación de pose** sitúa puntos o segmentos asociados al cuerpo. El **seguimiento** mantiene la correspondencia entre esas estimaciones en fotogramas sucesivos. Un programa puede detectar correctamente dos personas y, aun así, intercambiar sus identificadores cuando se cruzan. También puede seguir un cuerpo y situar mal un punto particular.

El aprendizaje profundo interviene cuando una red neuronal ajustada con ejemplos produce esas estimaciones. No convierte cada punto en una observación anatómica directa. La definición del

punto procede de las etiquetas, de la representación y del objetivo del sistema. Por eso una línea coloreada que une cadera y rodilla debe leerse como una construcción del análisis, antes de interpretarla como eje femoral o centro articular.

La distinción de Hall entre cinemática y cinética proporciona el marco para ordenar estas aplicaciones: describir posiciones y movimientos es una tarea diferente de explicar fuerzas y momentos. Los desarrollos computacionales amplían las formas de estimar las primeras, mientras que las segundas necesitan información y supuestos adicionales (Hall, 2015, pp. 48-49). Este artículo utiliza ejemplos matemáticos propios para mostrar dónde aparece cada decisión.

Del píxel al movimiento: cuatro niveles de información

Antes de interpretar un resultado hay que reconocer en qué nivel se encuentra. Una coordenada en píxeles, una distancia calibrada, un ángulo entre segmentos y una fuerza estimada no significan lo mismo. El paso entre niveles requiere operaciones explícitas; una imagen convincente no permite saltarse ninguna de ellas.

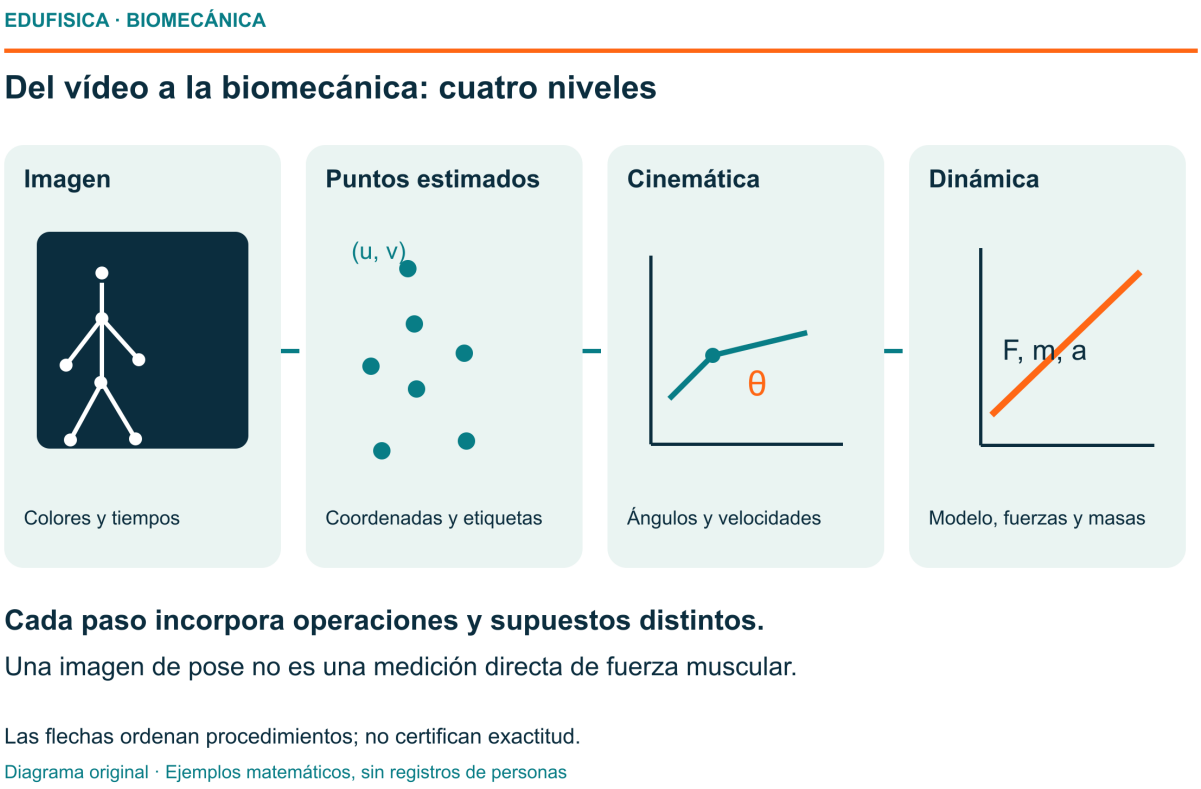
En el primer nivel están las **imágenes**: valores registrados por la cámara, junto con su resolución y sus tiempos. En el segundo aparecen los **puntos estimados**: pares de coordenadas, identificadores y, en algunos sistemas, indicadores de confianza. En el tercero se calculan **variables cinemáticas** mediante geometría, referencias espaciales y diferenciación temporal. En el cuarto se incorporan **modelos dinámicos** para relacionar el movimiento con fuerzas, masas o contactos.

Esta secuencia ayuda a describir un fallo sin atribuirlo a todo el sistema. Si el tobillo desaparece detrás de un banco, el problema comienza en la visibilidad. Si el tobillo se identifica correctamente pero la escala pertenece a otro plano, el fallo está en la conversión espacial. Si las posiciones son razonables y la aceleración presenta saltos, hay que revisar tiempos, ruido y procesamiento. Son preguntas diferentes, con soluciones posibles diferentes.

Por ejemplo, que un archivo contenga tres coordenadas por punto no demuestra que se haya medido profundidad con tres instrumentos. La tercera coordenada podría proceder de una reconstrucción multivista, de un modelo ajustado o de una predicción monocular. El informe debería decir cuál. La etiqueta «3D» describe una salida; para valorar su alcance también hace falta conocer cómo se obtuvo.

Información y operaciones necesarias

Nivel	Ejemplo de salida	Pregunta de control
Imagen	Fotograma con fecha temporal	¿Qué registró realmente la cámara?
Pose	Rodilla: coordenadas y etiqueta	¿Qué punto define esa etiqueta?
Cinemática	Ángulo, velocidad o trayectoria	¿Qué referencias, escala y tiempos se utilizaron?
Dinámica	Momento neto estimado	¿Qué fuerzas, parámetros y restricciones entraron en el modelo?



Cuatro representaciones originales distinguen imagen, puntos, cinemática y dinámica. Cada nivel conserva entradas distintas.
EDUFISICA · Diagrama didáctico original.

La cámara proyecta el espacio: perspectiva y profundidad

La imagen de una cámara es una proyección, no una copia del espacio. En un modelo ideal de cámara, un punto más lejano puede ocupar menos píxeles aunque conserve su tamaño físico. Esta relación explica por qué una escala colocada cerca del objetivo no describe automáticamente todo el campo.

Un modelo de cámara estenopeica, sin distorsión, permite formular la idea mediante $u=fX/Z+c_x$ y $v=fY/Z+c_y$. Las coordenadas X,Y,Z se expresan en el sistema de la cámara; f es la distancia focal expresada aquí en píxeles, y c_x,c_y localizan el centro principal de la imagen. No se trata de una calibración de un teléfono concreto, sino de una representación geométrica para razonar.

En un ejemplo hipotético, un segmento de 0,50 m paralelo al plano de imagen se observa a 5 m de profundidad con $f=1000$ píxeles. Su longitud proyectada es $1000(0{,}50)/5=100$ píxeles. Si se sitúa a 4 m, ocupa 125 píxeles; a 6 m, aproximadamente 83,33. La longitud física permanece igual. Utilizar la escala del primer plano para el segundo produciría una diferencia aparente que procede del montaje.

También importa la orientación. Un segmento que gira hacia el eje óptico puede acortar su proyección sin acortarse físicamente. Una comparación del mismo estudiante desde dos posiciones de cámara podría mostrar ángulos o distancias distintos por esta causa. La precaución no consiste en descartar el vídeo: consiste en formular una pregunta que la proyección permita responder.

Para una medición bidimensional, resulta más defendible estudiar un movimiento aproximadamente contenido en el plano calibrado y registrar cuándo se aparta de él. Hall sitúa la posición de la cámara y la naturaleza del movimiento entre las decisiones del análisis cinematográfico (2015, p. 48). La calibración y sus límites se desarrollan también en el videoanálisis con Kinovea.



Caso hipotético con f = 1000 px y longitud de 0,50 m. A profundidades de 4, 5 y 6 m, las proyecciones son 125, 100 y 83,33 px.
EDUFISICA · Diagrama didáctico original.

Estimación 2D y 3D: un mismo ángulo puede proyectarse de otra manera

Un ángulo calculado en la imagen relaciona dos proyecciones. Un ángulo tridimensional relaciona vectores espaciales. Ambos pueden ser útiles, pero describen operaciones distintas. Si una parte del movimiento sale del plano observado, la diferencia entre esos ángulos puede aumentar aunque los puntos de la imagen se localicen sin error.

La geometría permite demostrarlo sin atribuir valores a una persona. Tomemos dos vectores hipotéticos desde una articulación: $a=(1,0,1)$ y $b=(0,1,1)$. Su producto escalar es 1 y cada norma es $\sqrt{2}$. Por tanto, $\cos\theta=(a\cdot b)/(|a||b|)=1/2$, de modo que el ángulo espacial es 60°. Si la proyección ortográfica conserva solo las dos primeras coordenadas, los vectores se convierten en $(1,0)$ y $(0,1)$: el ángulo visible es 90°.

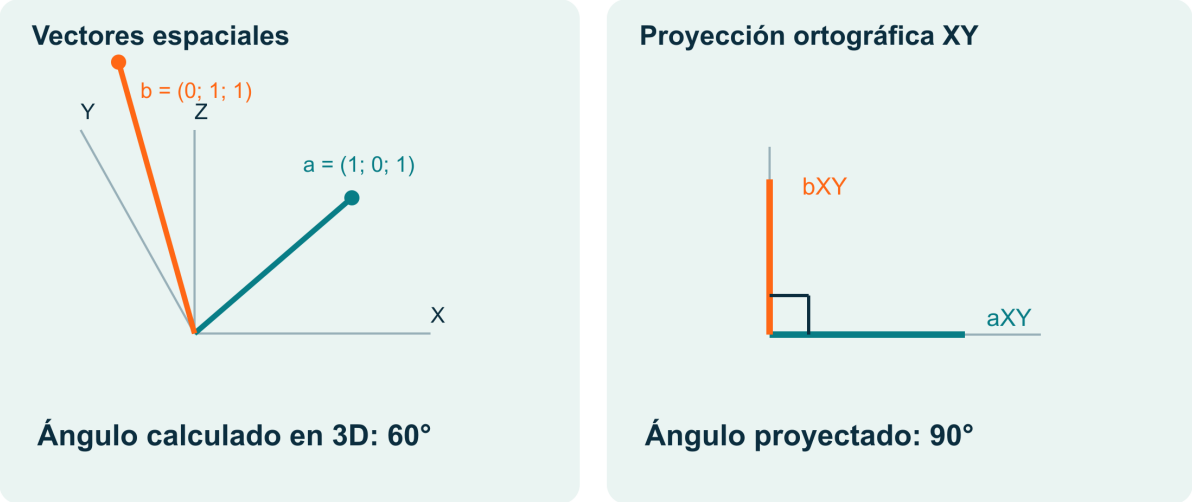
La diferencia de 30° no representa un error del detector en este ejemplo. Proviene de haber respondido una pregunta espacial con una proyección particular. Esta demostración tampoco reproduce una rodilla humana, una cámara en perspectiva ni un valor típico de error. Su finalidad es delimitar la operación. En una grabación real habría que añadir la geometría de cámara y las incertidumbres de los puntos.

Con una sola cámara, dos configuraciones espaciales pueden generar imágenes parecidas. Un método monocular que devuelve profundidad incorpora relaciones aprendidas o restricciones del modelo para resolver parte de esa ambigüedad. La solución obtenida puede ser compatible con la imagen y no corresponder con suficiente exactitud a una situación concreta. Esa posibilidad debe formar parte del diseño de la validación.

El nombre de la variable merece el mismo cuidado. «Ángulo entre muslo y pierna proyectados» informa mejor que «flexión exacta de rodilla» cuando no se ha definido una referencia anatómica tridimensional. Para comparar publicaciones, primero hay que comprobar que sus ángulos representan la misma convención y que la tarea y los planos son comparables.

EDUFISICA · BIOMECÁNICA

Un ángulo espacial y su proyección



Dibujo oblicuo ilustrativo a la izquierda; valores obtenidos por producto escalar.

Diagrama original · Ejemplos matemáticos, sin registros de personas

Modelo original: vectores (1; 0; 1) y (0; 1; 1), con 60° en 3D y 90° en proyección XY. No reproduce una articulación humana.
EDUFISICA · Diagrama didáctico original.

Aprendizaje profundo: qué ajustan las redes neuronales

Una red neuronal transforma sus entradas mediante operaciones parametrizadas. Durante el entrenamiento, esos parámetros se ajustan para reducir una función de error respecto de ejemplos y etiquetas. Durante el uso, el sistema aplica los parámetros obtenidos. Distinguir ambos momentos permite preguntar qué información sustenta realmente una estimación de pose.

Una operación elemental puede escribirse como $z=\sum_i w_i x_i + b$, seguida de una función de activación. Las entradas, los pesos y el sesgo adquieren significado dentro de una arquitectura. Las

redes convolucionales utilizan filtros sobre regiones de una imagen; las arquitecturas temporales incorporan relaciones entre muestras de una secuencia. El libro coordinado por Kumar y colaboradores presenta estas familias y sus diferencias de procesamiento (Kumar et al., 2026, pp. 187-191).

En pose, los ejemplos etiquetados especifican qué puntos debe localizar el sistema. Si una etiqueta corresponde a una región visible de la cadera, no basta con renombrarla como centro geométrico de la articulación. Un modelo posterior podría construir una estimación anatómica utilizando otras referencias, pero esa sería una transformación adicional. La documentación del conjunto de puntos debería conservar esa distinción.

El tamaño de la red no explica por sí solo su utilidad. Para el docente o investigador resulta más relevante saber si la salida representa la variable buscada, si puede revisarse y si se ha contrastado en condiciones pertinentes. Una predicción rápida de puntos superficiales puede servir para una descripción limitada y ser insuficiente para otra que dependa de profundidad, rotación axial o contacto.

Asimismo, una red ajustada no debe describirse como si aprendiera de cada estudiante al procesar una grabación. El ajuste con nuevos datos, cuando existe, requiere un procedimiento definido. La utilización ordinaria de un modelo y su reentrenamiento son operaciones diferentes. En un informe de movimiento debe indicarse qué versión y qué procedimiento produjeron la salida.

Oclusiones, etiquetas y seguimiento de varias personas

Una articulación puede quedar oculta por el propio cuerpo, un implemento o otra persona. El sistema podría omitir el punto, estimarlo a partir del contexto o conservar una trayectoria previa. Cada respuesta tiene consecuencias distintas para el análisis. Una línea continua en pantalla no demuestra una observación continua del punto.

Supongamos, de manera hipotética, que un balón tapa una muñeca durante tres fotogramas. Interpolar entre las posiciones visibles antes y después produce una trayectoria matemática. Puede ser útil para un intervalo delimitado, pero no recupera el movimiento que no fue registrado. Si durante ese intervalo se busca el máximo de velocidad de la mano, la interpolación podría alterar precisamente el resultado de interés.

Con varias personas aparece otra dificultad: asignar cada conjunto de puntos al participante correcto. La proximidad de cuerpos o un cruce puede comprometer la correspondencia temporal. Un error de identidad no se corrige promediando más fotogramas de la trayectoria equivocada. Debe revisarse el segmento donde cambia la asignación y conservar una señal de que ese tramo no es utilizable.

Los indicadores de confianza ayudan a organizar esa revisión, pero su interpretación depende del sistema. Un valor de 0,9 no puede convertirse sin comprobación en «90 % de probabilidad de centro articular correcto». Para una tarea propia puede definirse una regla de exclusión y estudiar su efecto; esa regla será parte del protocolo, no una ley general de la biomecánica.

En el ámbito escolar, una estrategia práctica es comenzar con un participante y un gesto sencillo, revisando varios fotogramas manualmente. Esta propuesta didáctica reduce la cantidad de decisiones simultáneas del ejercicio de análisis. No sustituye una validación ni promete una precisión determinada: facilita identificar qué información se conserva y cuál falta.

Fotogramas, frecuencia de captura y sincronización temporal

El análisis temporal depende de cuándo se capturó cada imagen. La frecuencia de reproducción, la frecuencia de captura y la velocidad de procesamiento del algoritmo son magnitudes diferentes. Confundirlas modifica duraciones y velocidades, incluso cuando las coordenadas espaciales son correctas. El archivo debe conservar tiempos o una frecuencia de captura verificable.

En una secuencia hipotética de 60 imágenes por segundo, 12 intervalos entre fotogramas equivalen a $12/60=0{,}20$ s. Ese tramo incluye 13 fotogramas si se cuentan ambos extremos. Si el vídeo se reproduce a 30 imágenes por segundo, puede verse más lento; la duración física del movimiento no cambia por esa reproducción. El cálculo debe utilizar los tiempos de captura.

Una tasa de procesamiento de 200 fotogramas por segundo describe cuánto trabajo computacional se completa bajo unas condiciones determinadas. No crea 200 observaciones originales por segundo a partir de una cámara de 30. Una interpolación puede producir puntos temporales adicionales, pero estos son estimaciones entre observaciones y deben identificarse como tales.

En una reconstrucción con varias cámaras, también importa que los fotogramas correspondan al mismo instante. Si un punto se desplaza a 2 m/s, una diferencia temporal hipotética de 0,02 s representa 0,04 m de desplazamiento durante ese intervalo. Esta multiplicación no predice el error tridimensional final: muestra por qué mezclar instantes distintos puede afectar la geometría utilizada para reconstruirlos.

La frecuencia apropiada depende de la duración del fenómeno que se busca describir. Si el contacto de interés dura unos pocos intervalos de la grabación, la identificación de inicio y fin tendrá una resolución temporal limitada. Es preferible informar ese límite y revisar la pregunta que presentar un decimal calculado como si procediera de una observación continua.

Calibrar la imagen antes de comparar distancias

Calibrar consiste en relacionar la representación registrada con referencias conocidas. En un análisis plano puede establecerse una escala local; en una reconstrucción espacial se necesitan parámetros de cámara y relaciones entre vistas. La elección depende de la pregunta. Una regla visible en pantalla no convierte toda la escena en un plano métrico.

Un ejemplo sencillo parte de una referencia coplanar de 1 m que ocupa 250 píxeles. La escala es $1/250=0{,}004$ m por píxel. Un desplazamiento de 75 píxeles en ese plano corresponde a 0,30 m bajo las condiciones de ese modelo. Si la referencia se midió como 240 píxeles, el mismo

desplazamiento produciría 0,3125 m. La diferencia de 0,0125 m muestra la dependencia respecto de la calibración.

Por tanto, deben registrarse el objeto de referencia, su dimensión, su posición y el procedimiento de selección. Si el sujeto cambia de profundidad, la escala local podría dejar de ser pertinente. Si la cámara se mueve o modifica el zoom, hay que revisar si la relación inicial continúa describiendo la secuencia. Estas decisiones forman parte de los datos del análisis.

Para un sistema espacial, los parámetros internos describen aspectos de la proyección y los externos relacionan posición y orientación de las cámaras con un sistema común. La distorsión de lente introduce otra transformación que debe considerarse cuando sea relevante. No corresponde atribuir un resultado a la red neuronal si el error procede de una relación geométrica anterior.

La calibración permite convertir o reconstruir; la validación pregunta si el procedimiento produce resultados adecuados frente a una referencia pertinente. Un sistema puede estar calibrado y mantener errores debidos a oclusiones, tiempos o definición de puntos. Utilizar ambos términos como sinónimos impide saber qué se ha comprobado y qué continúa siendo una hipótesis.

Triangulación multivista: reconstruir no significa observar directamente

Varias cámaras calibradas permiten relacionar proyecciones del mismo punto. La triangulación busca una posición espacial compatible con esas vistas y sus parámetros. En datos reales, las líneas geométricas pueden no coincidir exactamente; el procedimiento debe resolver esa discrepancia. Su resultado depende de las correspondencias, los tiempos y la calibración utilizada.

Una representación estereoscópica simplificada explica la relación entre profundidad y disparidad. Para dos cámaras ideales paralelas con separación B, distancia focal f y disparidad horizontal d, se obtiene $Z=fB/d$. La disparidad es la diferencia entre posiciones correspondientes en ambas imágenes después de establecer una geometría común. Esta fórmula no describe automáticamente cualquier disposición de teléfonos.

En un ejemplo hipotético, $f=800$ píxeles, $B=0{,}50$ m y $d=100$ píxeles dan $Z=4$ m. Si la disparidad utilizada fuera 95 píxeles, la profundidad sería aproximadamente 4,21 m; con 105 píxeles, aproximadamente 3,81 m. No son mediciones de un sistema comercial. Son resultados del modelo elegido que hacen visible cómo una diferencia en la correspondencia modifica la reconstrucción.

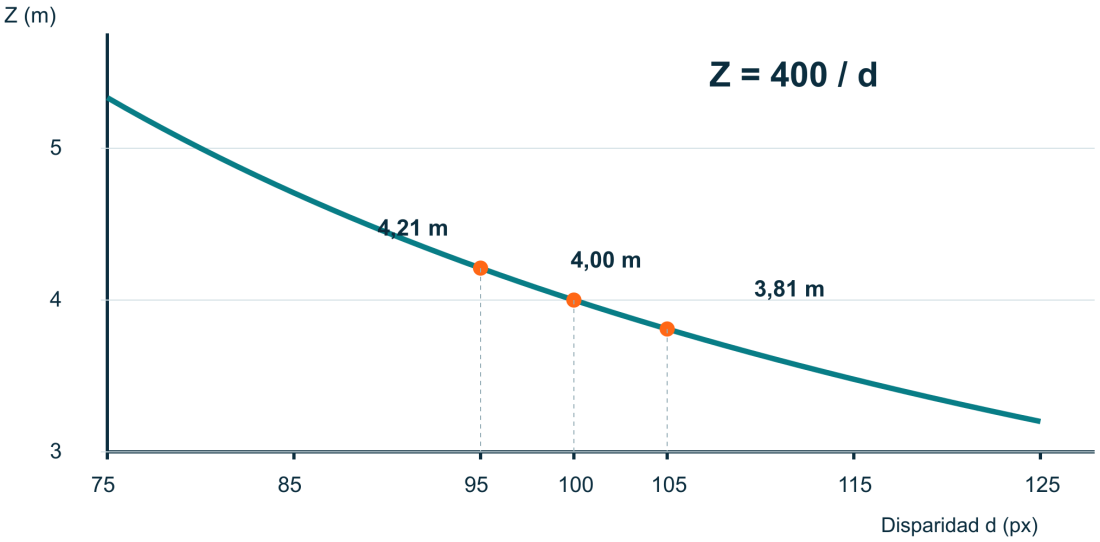
La sensibilidad local puede expresarse como $(\partial Z)/(\partial d)=-(fB)/(d^2)$. En este caso vale −0,04 m por píxel en torno a la disparidad de 100. Para un cambio pequeño, proporciona una aproximación; para los cambios de cinco píxeles anteriores, el cálculo exacto permite ver que las variaciones no son perfectamente simétricas. Esta comparación enseña a distinguir una derivada local de una equivalencia válida para cualquier rango.

En un protocolo multivista hay que comprobar que todas las cámaras incluyen la zona relevante, que los puntos corresponden al mismo participante y que los tiempos representan la misma fase del gesto. Añadir cámaras modifica las oportunidades geométricas, pero también añade relaciones que deben verificarse. La cantidad de dispositivos no sustituye la calidad de esas relaciones.

EDUFISICA · BIOMECÁNICA

Disparidad y profundidad reconstruida

Cámaras ideales paralelas: f = 800 px; B = 0,50 m



Modelo hipotético: cinco píxeles de diferencia no producen variaciones simétricas.

Diagrama original · Ejemplos matemáticos, sin registros de personas

Modelo hipotético Z = 400/d, con d en píxeles y Z en metros. Las disparidades de 95, 100 y 105 dan profundidades de 4,21, 4 y 3,81 m.
EDUFISICA · Diagrama didáctico original.

Calcular ángulos: referencias, convenciones y centros articulares

Un ángulo necesita una definición antes de convertirse en resultado. Puede relacionar segmentos, describir una orientación respecto del laboratorio o formar parte de una secuencia tridimensional de rotaciones. Las respuestas no son intercambiables. La convención elegida debe mantenerse al comparar participantes, sesiones y publicaciones sobre el mismo movimiento.

Para tres puntos planos A, B y C, con B como vértice, pueden construirse $a=A-B$ y $b=C-B$, y calcular $\theta=\arccos[(a\cdot b)/(|a||b|)]$. El resultado de esta operación es un ángulo entre dos vectores. Si B se denomina rodilla, A cadera y C tobillo, todavía hay que explicar qué representan sus coordenadas y cómo se relaciona el ángulo geométrico con la convención de flexión.

En un ejemplo de vectores alineados en sentidos opuestos, el ángulo incluido es 180°. Una convención de flexión que utiliza la extensión como cero puede expresar $180°-\theta$. Aplicar esa resta a todos los movimientos sin describir el plano o la referencia sería insuficiente. Las rotaciones fuera del plano y las

diferencias entre puntos estimados y referencias anatómicas requieren una interpretación más específica.

Además, un punto articular no define por sí solo la orientación completa de un segmento en el espacio. Una línea entre dos puntos determina su dirección longitudinal, pero deja abierta la rotación alrededor de esa línea. Para analizar esa rotación hacen falta referencias adicionales o restricciones explícitas. Dibujar un cilindro alrededor de la línea no añade la información que falta.

Para la enseñanza, puede ser útil comenzar con ángulos planos entre segmentos abstractos y, después, contrastarlos con una grabación. La progresión permite que el alumnado identifique la operación antes de asociarla a una articulación. El objetivo del ejercicio no es clasificar cuerpos por un ángulo aislado, sino comprender cómo se construye y qué condiciones limitan su interpretación.

Exactitud, repetibilidad y validación de la estimación de pose

Un sistema puede repetir una estimación y mantener una diferencia respecto de la referencia. También puede presentar un error medio pequeño junto con errores grandes en algunos fotogramas. Por eso la evaluación debe describir varias propiedades y conservar su contexto. Una cifra resumida no explica dónde ni cuándo aparece la discrepancia.

El error medio absoluto se calcula como $\mathrm{MAE}=(1/n)\sum_i|\hat{x}_i-x_i|$. La raíz del error cuadrático medio es $\mathrm{RMSE}=\sqrt{((1/n)\sum_i(\hat{x}_i-x_i)^2)}$. La diferencia media con signo informa de sesgo bajo esa definición. Son operaciones distintas sobre las mismas discrepancias y sus unidades deben corresponder a la variable estudiada.

Tomemos errores hipotéticos de posición −2, 1, 3 y −2 cm. El sesgo medio es cero, el MAE es 2 cm y el RMSE es aproximadamente 2,12 cm. Que los signos se compensen no significa que los puntos se hayan localizado sin error. Este ejemplo explica por qué una media con signo no puede utilizarse como sustituto de una medida de magnitud.

La comparación debe realizarse entre variables equivalentes y muestras sincronizadas. Si el sistema estima un punto visible y la referencia utiliza otro definido anatómicamente, parte de la discrepancia procede de esa definición. Si ambos describen eventos distintos, el error temporal puede mezclarse con el espacial. Antes de interpretar la cifra hay que revisar esa correspondencia.

La repetibilidad exige repetir el procedimiento bajo condiciones definidas. Una revisión visual de una sola grabación no la demuestra. La validez para una tarea tampoco se extiende automáticamente a otras edades, velocidades o movimientos. La pregunta operativa es si el error y la incertidumbre del procedimiento permiten distinguir el fenómeno que se pretende estudiar.

De posiciones a velocidades: cómo se amplifica el ruido

La diferenciación temporal convierte cambios de posición en velocidad y cambios de velocidad en aceleración. Si las posiciones contienen pequeñas oscilaciones, estas operaciones pueden amplificarlas. El problema adquiere especial importancia cuando se interpreta un máximo breve. Una trayectoria visualmente aceptable no basta para valorar la calidad de sus derivadas.

Con intervalos regulares puede aproximarse la velocidad mediante $v_i=(x_{i+1}-x_{i-1})/(2\Delta t)$. En una demostración hipotética, el punto permanece físicamente quieto pero los errores en las dos muestras vecinas son +0,005 y −0,005 m. Con $\Delta t=0{,}01$ s, la operación produce una velocidad de magnitud 0,50

m/s. Ese resultado procede del error elegido, no de un desplazamiento real.

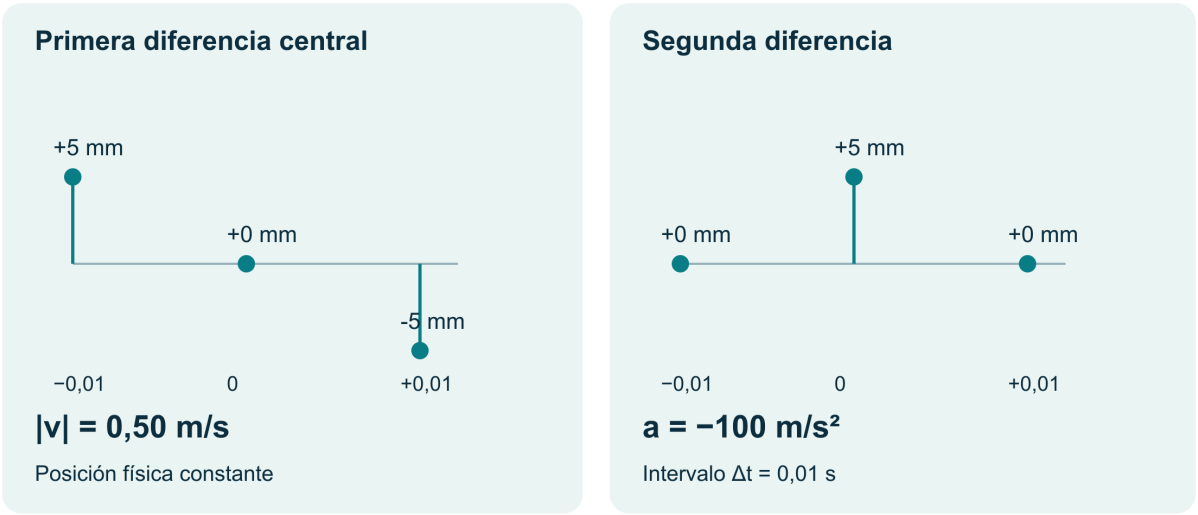
La segunda diferencia, $a_i=(x_{i+1}-2x_i+x_{i-1})/(\Delta t)^2$, permite mostrar un efecto todavía mayor. Si las dos muestras vecinas tienen error cero y la central un error de +0,005 m, con el mismo intervalo aparece −100 m/s². No es una predicción de error habitual de una cámara. Es un cálculo exacto bajo una secuencia hipotética que explica por qué la aceleración necesita un tratamiento cuidadoso.

Filtrar puede reducir determinadas oscilaciones, pero también modificar cambios rápidos y extremos. Interpolar una oclusión produce datos dependientes del método. Cada operación debería quedar registrada con sus parámetros y su posición en la secuencia de procesamiento. La comparación entre una señal filtrada y otra sin filtrar exige reconocer que se han construido de manera distinta.

Una comprobación útil consiste en revisar posiciones, tiempos y derivadas conjuntamente. Un pico de aceleración coincidente con un cambio de identidad o una oclusión merece inspección antes de atribuirse a una acción muscular. Para ampliar esta relación entre señal y magnitud puede consultarse el análisis cinemático con Tracker.

EDUFISICA · BIOMECÁNICA

Pequeño error de posición, grandes derivadas



Dos secuencias de error distintas; estos valores no son errores medidos de un sistema.

Diagrama original · Ejemplos matemáticos, sin registros de personas

Dos secuencias hipotéticas de errores milimétricos, con intervalo de 0,01 s, producen magnitud de velocidad de 0,50 m/s y aceleración de −100 m/s². No son registros reales.
EDUFISICA · Diagrama didáctico original.

Datos de entrenamiento y prueba: comprobar la generalización

Evaluar un estimador requiere separar el ajuste del modelo de su contraste. Si se prueba únicamente sobre ejemplos utilizados para ajustar parámetros, no se demuestra su desempeño en datos nuevos. La separación también debe considerar quién aparece en las secuencias y cómo se parecen entre sí, además del número de imágenes.

En una situación hipotética, distribuir fotogramas alternos de una misma grabación entre entrenamiento y prueba puede colocar imágenes casi idénticas en ambos grupos. El resultado responde a una pregunta diferente de probar con participantes y situaciones independientes. Para estudiar el funcionamiento en nuevas personas, una separación por participante proporciona una condición más pertinente que una separación aleatoria de fotogramas vecinos.

Este principio no obliga a que todos los estudios respondan la misma pregunta. Una investigación puede buscar desempeño dentro de una tarea muy delimitada; otra puede estudiar transferencia entre escenarios. Lo necesario es que el diseño y la interpretación coincidan. El número total de fotogramas no informa por sí solo de la diversidad de cuerpos, acciones y condiciones representadas.

También hay que distinguir entrenamiento del estimador, selección de parámetros del procesamiento y evaluación final. Elegir un filtro después de observar el resultado de prueba modifica el procedimiento que se está contrastando. Una comparación reproducible registra qué decisiones se fijaron antes de evaluar y cuáles se modificaron al revisar fallos.

Para un docente que utiliza una herramienta ya ajustada, el equivalente práctico es comprobar la documentación y contrastar una muestra de la tarea propia. No necesita entrenar una red para formular preguntas rigurosas: qué población se estudió, qué movimiento se registró, qué referencia se utilizó y qué resultado se validó. Las respuestas delimitan el uso posible.

Qué muestran las investigaciones sobre análisis sin marcadores

Las investigaciones deben leerse junto con su procedimiento y su muestra. Un resultado obtenido con varias cámaras y modelado físico no describe cualquier aplicación monocular. Del mismo modo, un error en una tarea concreta no resume todas las formas de captura. Comparar diseños ayuda a ubicar cada evidencia sin convertirla en una promesa general.

Uhlrich y colaboradores integraron vídeos de dos o más teléfonos, estimación de puntos, reconstrucción, modelado y simulación en OpenCap. La validación de laboratorio incluyó diez personas sanas y tareas como caminar, sentarse y levantarse, sentadillas y saltos. El error angular medio absoluto comunicado fue 4,5°. Los autores identificaron límites de generalización y señalaron que habían utilizado personas sanas para reproducir modificaciones de movimiento. Ese diseño no equivale a validar diagnósticos en todas las poblaciones (Uhlrich et al., 2023).

Rode y colaboradores compararon once estimadores monoculares con una referencia óptica utilizando ejercicios de veinticinco participantes sin alteraciones que impidieran realizarlos. Su conjunto Physio2.2M reunió 2,2 millones de fotogramas. El estudio encontró diferencias entre métodos y dimensiones, y destacó las limitaciones de profundidad de las soluciones monoculares. Sus resultados corresponden a esas tareas y a esa muestra; no establecen un error universal para cualquier cámara o ejercicio (Rode et al., 2025).

La lectura comparada orienta una decisión metodológica: antes de elegir un sistema, especificar qué variable se necesita y qué condiciones deben representar sus comprobaciones. Para una aplicación nueva resulta insuficiente copiar una cifra de precisión y colocarla junto al nombre de la herramienta. La relación entre esa cifra, el montaje y la población es parte de la evidencia.

Por qué una pose no mide directamente la fuerza muscular

La pose describe una configuración estimada del cuerpo. Para relacionarla con fuerzas se necesita un modelo dinámico, parámetros y condiciones de interacción. Diferentes distribuciones de fuerza pueden ser compatibles con un movimiento parecido. Por tanto, un gráfico muscular añadido al vídeo debe identificarse según el procedimiento que lo produjo.

Las ecuaciones de Newton relacionan fuerza resultante, masa y aceleración; el análisis rotacional requiere además momentos, propiedades inerciales y referencias. En un sistema segmentario importan fuerzas externas, contactos y conexiones entre partes. Conocer posiciones por sí solo no proporciona todas esas entradas. Una estimación de contacto a partir de un modelo es una entrada construida y no debe llamarse medición de plataforma de fuerza.

Incluso cuando se calcula un momento articular neto, repartirlo entre músculos es otro problema. Si dos músculos tienen posibilidades de producir momentos en el mismo sentido, diferentes combinaciones de fuerza pueden satisfacer una ecuación neta. Añadir restricciones o un criterio de optimización permite obtener una solución del modelo, pero no transforma esa solución en una observación individual de cada músculo.

Por ejemplo, la igualdad hipotética $10 = 0{,}04F_1 + 0{,}02F_2$ admite $F_1 = 250$ N y $F_2 = 0$, o $F_1 = 200$ N y $F_2 = 100$ N, entre otras combinaciones. Es una demostración algebraica sin anatomía individual: muestra que una ecuación puede dejar varias incógnitas abiertas. Un modelo más detallado incorpora información, pero también debe declarar sus supuestos.

Esta separación enlaza el análisis de pose con la interpretación del torque muscular y con el modelado musculoesquelético. Para comunicar resultados conviene utilizar expresiones como «posición estimada», «momento neto calculado» o «fuerza muscular predicha por el modelo», según corresponda. Las palabras preservan la procedencia de la magnitud.

Diseñar un protocolo de captura con una pregunta concreta

Un protocolo útil empieza por la variable que se quiere estudiar y termina en un resultado revisable. Entre ambos extremos deben quedar definidos tarea, participantes, referencias y procesamiento. Empezar por activar todas las funciones de una aplicación puede producir muchas cifras y pocas respuestas defendibles sobre el movimiento observado.

La siguiente propuesta original sirve para una actividad de análisis educativo. Primero se formula una pregunta delimitada, por ejemplo, describir el ángulo proyectado entre dos segmentos en una fase de un gesto lento. Después se establece la fase, los puntos, el plano esperado y el criterio de identificación. La pregunta evita atribuir fuerzas internas o diferencias clínicas a una descripción plana.

Se registra el montaje antes de grabar: posición de cámara, referencia de escala si se necesitan distancias, resolución, frecuencia temporal y condiciones de visibilidad. Una prueba breve permite revisar que el movimiento relevante permanece en el encuadre y que los puntos no se ocultan durante la fase analizada. Si

la pregunta exige profundidad, se reconsidera el montaje y el método.

El procesamiento conserva una copia del registro original y una secuencia de decisiones. Se documentan versión del estimador, conjunto de puntos, tramos omitidos, interpolaciones y filtros. La revisión combina el vídeo con las coordenadas y las variables derivadas. Un resultado que no permite localizar el fotograma o el cálculo correspondiente dificulta distinguir interpretación de procesamiento.

Finalmente se informa qué puede concluirse y qué no se estudió. Una comparación educativa puede describir una diferencia observada en dos ejecuciones concretas; no demuestra que un método de enseñanza cause una mejora general. La organización del protocolo permite ampliar posteriormente la pregunta sin confundir la actividad inicial con un estudio de eficacia.

Registro mínimo propuesto para una tarea propia

Decisión	Información a conservar
Pregunta	Variable, fase y comparación prevista
Montaje	Cámara, plano, referencias y tiempos
Estimación	Versión, puntos y correspondencias
Procesamiento	Exclusiones, filtros e interpolaciones
Interpretación	Alcance del resultado y condiciones que lo limitan

Un ejemplo docente: comparar dos registros sin confundir cuerpo y cámara

Una actividad de aula puede utilizar la estimación de pose para enseñar a construir preguntas, además de mostrar un gesto. El siguiente caso es hipotético y no presenta resultados de estudiantes reales. Su finalidad es contrastar explicaciones posibles de una diferencia visual y decidir qué información falta para sostener cada una.

Dos grabaciones muestran el mismo gesto en un plano aproximado. En la primera, una referencia de 1 m ocupa 250 píxeles; en la segunda, 200. Un punto recorre 75 píxeles en cada secuencia durante 0,20 s. Con sus escalas correspondientes, los desplazamientos son 0,30 y 0,375 m, y las velocidades medias calculadas son 1,50 y 1,875 m/s.

Esas operaciones son correctas para las escalas y los tiempos elegidos, pero todavía no explican por qué cambian los resultados. Podrían corresponder a desplazamientos físicos diferentes, a referencias situadas en otro plano o a un movimiento fuera de la zona calibrada. El ejercicio consiste en formular esas posibilidades y comprobarlas en el montaje, en lugar de atribuir de inmediato la diferencia a una mejor ejecución.

Después se introduce una versión en la que ambos desplazamientos físicos se fijan matemáticamente en 0,30 m. Si una cámara los representa con 75 píxeles y otra con 60, las escalas de 0,004 y 0,005 m por píxel recuperan el mismo valor. La demostración permite separar cantidad de píxeles de distancia física. No requiere clasificar al participante ni establecer una técnica ideal.

La ficha de trabajo puede pedir cuatro respuestas: qué se observó, qué se calculó, qué se supuso y qué se necesita comprobar. Estas categorías funcionan como guía de discusión de este ejercicio, no como instrumento validado de evaluación. El docente puede adaptar lenguaje y complejidad sin convertir la aplicación en una autoridad que decida la nota.

Una extensión consiste en comparar la selección manual de tres fotogramas con la estimación automática. Las discrepancias se estudian sobre puntos definidos, no como una competición entre «humano» y «máquina». Ambos procedimientos pueden presentar limitaciones diferentes. La pregunta relevante es cómo esas discrepancias afectan la variable que el grupo intenta comprender.

Cuándo resulta pertinente y cuándo hay que cambiar de método

La pertinencia de un método depende de la relación entre la pregunta y sus límites. Una descripción visual de una secuencia, una medición angular y una decisión clínica requieren evidencias distintas. El criterio de elección debe formularse antes de interpretar el resultado, incluyendo qué discrepancia sería relevante para la comparación prevista.

Puede resultar pertinente una tarea plana sencilla con puntos visibles, tiempos conservados y revisión de la salida. Si se pretende observar una tendencia amplia de una trayectoria, la pregunta no exige necesariamente el mismo detalle que identificar un pico breve. Cada caso debe indicar qué parte de la variable se busca y qué condiciones se aceptan.

Hay razones para cambiar de método cuando el fenómeno depende de profundidad no contrastada, rotación axial sin referencias suficientes, oclusiones prolongadas o eventos más breves que la resolución temporal. También cuando la herramienta no permite revisar puntos y procesamiento. En esas situaciones, producir más gráficos de la misma información no resuelve la falta que afecta a la pregunta.

Un criterio hipotético ayuda a concretarlo: si una comparación pretende distinguir diferencias de 2° y el procedimiento presenta discrepancias de magnitud semejante o superior bajo condiciones pertinentes, esa comparación necesita revisión. El número 2° no es un umbral universal ni una recomendación clínica; representa una pregunta elegida para mostrar que el error debe valorarse en relación con el efecto que se quiere distinguir.

Las decisiones sobre dolor, lesión, retorno al deporte o tratamiento incorporan información clínica que una pose aislada no proporciona. Una herramienta puede participar en una evaluación delimitada, pero su salida no reemplaza ese proceso. En Educación Física, describir aprendizaje y facilitar reflexión sobre el movimiento ofrece una aplicación diferente de diagnosticar condiciones a partir de un vídeo.

Grabaciones escolares: identidad, acceso y conservación

Una grabación corporal puede identificar a una persona aunque el análisis se centre en coordenadas. Antes de utilizarla conviene acordar para qué se registra, quién accede y qué ocurre con el archivo después. Estas decisiones pertenecen al diseño de la actividad, junto con las referencias espaciales y los criterios de análisis.

La propuesta educativa puede reducir información que no necesita: encuadrar la tarea, evitar participantes ajenos y utilizar identificadores de trabajo en los archivos. Un esqueleto superpuesto no convierte automáticamente un vídeo identificable en anónimo. Si se conserva rostro, voz, uniforme u otras referencias, esa identidad sigue formando parte del material.

También hay que distinguir procesamiento local y envío a una plataforma. El lugar donde se calculan los puntos modifica quién podría acceder a la grabación y qué condiciones rigen su conservación. Antes de utilizar una herramienta con alumnado se deben comprobar sus condiciones y las reglas institucionales aplicables. Esta comprobación no puede reemplazarse por una afirmación genérica de que el análisis «solo usa datos».

Para una demostración de geometría pueden utilizarse coordenadas sintéticas como las de este artículo. Permiten estudiar escalas, proyecciones y errores sin presentar movimientos inventados como si fueran registros personales. Cuando la pregunta requiere una ejecución real, su registro y utilización deben organizarse con las autorizaciones y medidas pertinentes al contexto.

El criterio editorial de una explicación científica es conservar únicamente la información necesaria para que el resultado pueda entenderse y revisarse. Publicar un vídeo completo no se deriva automáticamente de haber autorizado su uso en clase. Análisis, conservación y difusión son decisiones distintas que conviene resolver por separado.

Preguntas frecuentes sobre visión artificial y estimación de pose

Las dudas más útiles delimitan qué produce el sistema y qué pasos hacen falta para interpretarlo. No existe una respuesta única sobre precisión para todas las cámaras, tareas y poblaciones. Las siguientes respuestas conservan esa relación entre variable, procedimiento y alcance, de modo que puedan utilizarse al revisar una herramienta concreta.

¿Qué es la estimación de pose humana?

Es un procedimiento que sitúa puntos o segmentos asociados al cuerpo a partir de imágenes u otros datos. En vídeo, sus coordenadas son estimaciones definidas por el modelo y sus etiquetas. Para utilizarlas en biomecánica hay que conocer qué representan, cómo se relacionan con las referencias del cuerpo y qué variable se calcula después.

¿Una sola cámara puede medir el movimiento en tres dimensiones?

Una cámara registra una proyección del espacio. Un sistema monocular puede devolver una reconstrucción tridimensional mediante relaciones aprendidas o restricciones, pero esa salida incorpora estimación de profundidad. Su utilidad requiere contrastar el procedimiento en tareas pertinentes. La etiqueta 3D no demuestra que todas las coordenadas se hayan observado directamente ni que describan cualquier movimiento con el mismo error.

¿La estimación de pose mide la fuerza muscular?

La pose por sí sola no mide directamente la fuerza muscular. Para relacionar movimiento y fuerzas se necesitan modelos, parámetros y condiciones de interacción. Un momento articular neto y una fuerza muscular individual son resultados diferentes. Cuando un programa presenta fuerzas, el informe debe explicar cómo se calcularon o predijeron y qué información entró en ese procedimiento.

¿Más fotogramas por segundo significan mayor exactitud?

La frecuencia de captura describe la resolución temporal disponible, pero no resume la calidad de una medición. También importan visibilidad, exposición, calibración, definición de puntos y procesamiento. La velocidad de ejecución del algoritmo es otra magnitud. Procesar más rápido no añade observaciones originales a una grabación capturada con una frecuencia menor.

¿Cómo puede utilizarse esta tecnología en Educación Física?

Puede formar parte de actividades delimitadas de observación, geometría y reflexión sobre el movimiento. Una propuesta docente debe definir pregunta, puntos, fase, referencias y revisión de resultados. Las coordenadas estimadas no deberían convertirse sin contraste en diagnóstico, técnica ideal o calificación automática. La tarea puede enseñar a distinguir observación, cálculo y supuesto antes de ampliar su complejidad.

Referencias bibliográficas

1. Hall, S. J. (2015). *Basic biomechanics* (7th ed.). McGraw-Hill Education. Registro bibliográfico.

2. Kumar, A., Kumar, T. A., Ahuja, S., Ananth, J. P., & Manoj, S. O. (Eds.). (2026). *Physiotherapy using artificial intelligence: Enhancing biomechanics for optimal rehabilitation*. Wiley/Scrivener. ISBN 9781394391530.

3. Rode, D., Dunkel, A., Willi, R., Wolf, P., Xiloyannis, M., & Riener, R. (2025). Assessment of monocular human pose estimation models for clinical movement analysis. *Scientific Reports, 15*, Article 38767. https://doi.org/10.1038/s41598-025-22626-7

4. Uhlrich, S. D., Falisse, A., Kidziński, Ł., Muccini, J., Ko, M., Chaudhari, A. S., Hicks, J. L., & Delp, S. L. (2023). OpenCap: Human movement dynamics from smartphone videos. *PLOS Computational Biology, 19*(10), Article e1011462. https://doi.org/10.1371/journal.pcbi.1011462