Escuela Politécnica Superior

Estudio del estado del arte de los métodos de estimación de la pose humana en 3D

Grado en Ingeniería Informática

Trabajo Fin de Grado

Autor:
Joan Reig Doménech

Tutores:
Andrés Fuster Guilló
Jorge Azorín López

Universitat d'Alacant
Universidad de Alicante

Junio 2018

Índice

1. Introducción

1.1. Motivación

El presente Trabajo Fin de Grado está relacionado con el proyecto de investigación *Tech4Diet* llevado a cabo por el Departamento de Tecnología Informática y Computación donde Jorge Azorín y Andrés Fuster forman parte del equipo de investigación. Con anterioridad ya tuve vinculación con dicho proyecto al haber realizado las asignaturas de *Practicas Externas I y II* (34069 y 34070) en este departamento. Asimismo, este trabajo guarda relación con asignaturas del itinerario de Computación como *Visión Artificial y Robótica* (34030) o *Razonamiento Automático* (34031), entre otras.

Tech4Diet es un proyecto que desarrolla un sistema para medir la evolución física del cuerpo humano en determinados tratamientos haciendo uso de las tecnologías de visión artificial. El sistema utiliza imágenes 3D obtenidas a lo largo del tiempo, incluyendo así una cuarta dimensión. El resultado es un análisis preciso que resulta de gran utilidad tanto en tratamientos de adelgazamiento como en desarrollo muscular.

El sistema funcionaría de la siguiente manera: utilizando una sola cámara RGB-D de bajo coste (Kinect), se captarían los datos de profundidad más las imágenes de color. Al tomar las imágenes será el paciente el que se vaya moviendo sobre sí mismo para poder captar todo el cuerpo. Al mismo tiempo, se extraería la pose de la persona, para posteriormente poder registrar la nube de puntos obtenida. De esa nube de puntos se formaría un modelo en 3D al cual se le aplicarían texturas y se introduciría en un entorno virtual. Con más de un modelo se realizaría una animación donde se podría ver la deformación del cuerpo a lo largo del tiempo.

1.2. Objetivos

Uno de los desafíos del proyecto lo constituye el hecho de que la pose obtenida no es correcta en muchos casos, debido a que el algoritmo proporcionado por la Kinect no calcula correctamente el esqueleto de la persona cuando ésta se encuentra de perfil o de espaldas a la cámara.

El objetivo principal de este trabajo es realizar un estudio del estado del arte de los métodos de la estimación en tres dimensiones de la pose humana, para encontrar uno o más métodos que solucionen el problema y se adecuen al proyecto de investigación.

Para ello se seguirán una serie de pasos o subobjetivos que detallamos a continuación:

- Primero, recopilar información sobre los distintos métodos de estimación de la pose, ya que actualmente no se ha encontrado ningún trabajo que incluya una recopilación completa de los métodos más destacados de este ámbito.
- En segundo lugar, analizar las características y los problemas que presentan dichos métodos, para poder determinar en qué ámbitos o aplicaciones resultan más beneficiosos.
- Por último, comparar los métodos para conocer el rendimiento actual de los distintos algoritmos y así poder elegir el método que mejor se adapte y mejores resultados obtenga al incorporarlo en el proyecto de investigación.

1.3. Contexto

Cuando hablamos de estimación de la pose humana nos referimos a la tarea de capturar y analizar las articulaciones y el movimiento del cuerpo mediante técnicas de visión por computador, utilizando una imagen o una secuencia de imágenes para estimar la configuración de las partes del cuerpo humano. Esta área de investigación está en auge debido a sus muchas aplicaciones y a su gran complejidad. Aunque haya que superar problemas complejos, como la auto-oclusión de partes del cuerpo, hacerlo supone un desafío desde el punto de vista académico.

Atendiendo a la perspectiva práctica (*cf.* Fig. 1), las soluciones aportadas por la visión por computador son muy atractivas, dado que proporcionan métodos no invasivos para muchas aplicaciones: la comunicación humano-ordenador o humano-robot se puede facilitar si se pueden reconocer gestos que sirvan de comandos (Kondoři, 2014). En el ámbito de los videojuegos encontramos un gran uso de esta tecnología, no solo en forma de controles para el usuario a la hora de jugar (Suma *et al.*, 2011; Wouterse, 2015), sino también como herramientas para generar avatares en 3D (Condell, Moore and Moore, 2006). En el campo deportivo o médico lo visualizamos en herramientas de análisis y de estudio del comportamiento anatómico del cuerpo humano (Ortiz-catalan *et al.*, 2014). Asimismo, se han llevado a cabo investigaciones acerca del entendimiento del comportamiento y las relaciones humanas, cuya aplicación puede ayudar en las tareas de videovigilancia (Bhaltilak, Kaur and Khosla, 2014).

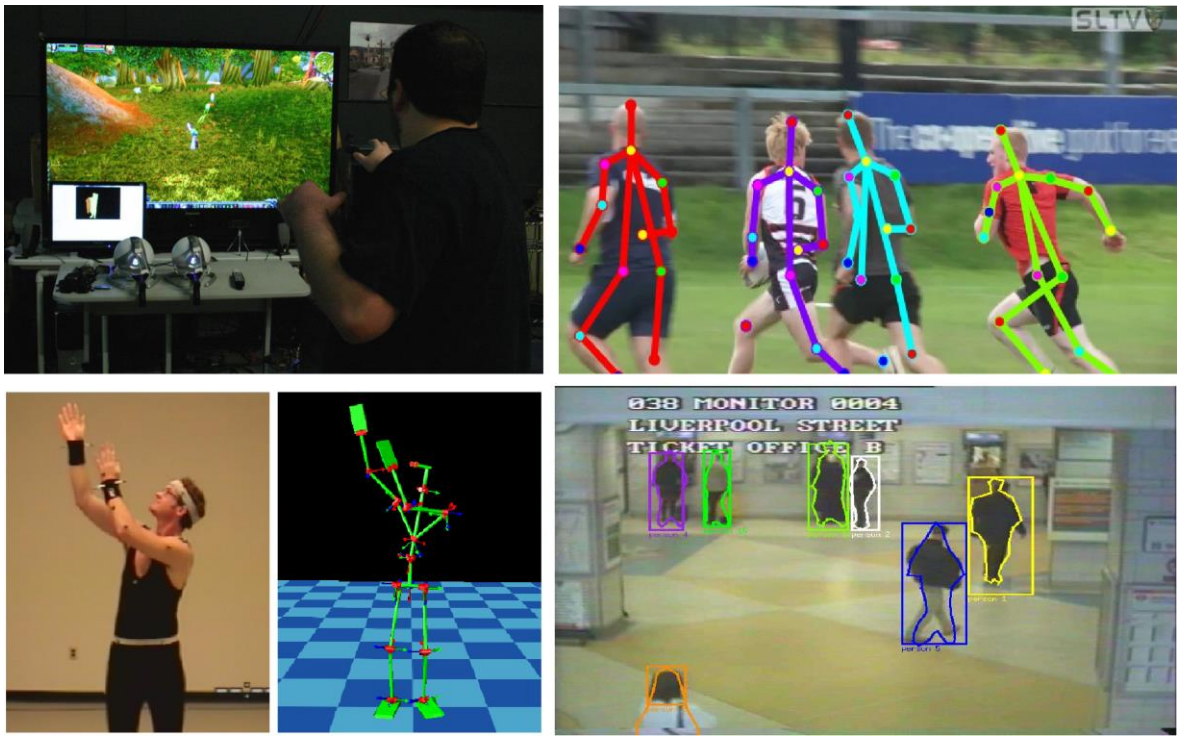

Fig. 1. ***Ejemplos de aplicaciones de la estimación de la pose.*** *De izquierda a derecha y de arriba abajo: control de videojuegos, deportes, creación de animaciones para videojuegos, videovigilancia (Fuente: Suma et al.* (2011)*, Iqbal, Milan y Gall* (2017)*, Zheng y Yamane* (2015)*, Satpathy, Siebel y Rodríguez* (2004)*).*

Para iniciarnos en este campo, los estudios realizados por Gavrila (1999) sobre los métodos iniciales en el campo de la visión por computador podemos considerarlos como un buen punto de partida. Por su parte, los trabajos de Moeslund *et al.* (2006) y Poppe (2007) cubren en gran medida el campo de los métodos de visión por computador usados para capturar el movimiento humano, aunque no se especializan en la estimación en 3D y se encuentran un tanto desactualizados. El estudio realizado por Sarafianos *et al.* (2016) incluye métodos más actuales, aunque deja de lado aquellos que utilizan información de profundidad como datos de entrada. Existen publicación centradas en tareas más específicas, como diseñar interfaces para comunicarse con los ordenadores mediante gestos (Wigdor and Wixon, 2011) o métodos de análisis del movimiento humano (Ye *et al.*, 2013), que hacen uso de métodos de estimación de la pose.

1.3.1. Forma y clasificación de los métodos de estimación de la pose 3D

En la Fig. 2 se pueden ver representados los pasos generales que presentan muchos de los métodos utilizados para la estimación de la pose en 3D, los cuales son:

1. Utilización de un modelo predefinido del cuerpo humano, el cual determinará si el método es *basado en modelo* (generativo), *libre de modelo* (discriminativo) o híbrido.
2. Utilización de información en 2D o 3D como fuente de información y/o medida de precisión.
3. Utilización de técnicas de preprocesado, como la extracción del fondo de la imagen o la eliminación de ruido.
4. Extracción y selección de características clave del cuerpo.
5. Obtención de una pose inicial en 3D.
6. Aplicación de modelos y restricciones para descartar poses no realistas.
7. Obtención del modelo final.

No resulta útil entrar en detalle sobre los pasos de los distintos métodos, puesto que la comunidad científica no se ha puesto de acuerdo en una taxonomía bien estructurada. Por consiguiente, aunque muchos métodos comparten pasos, otros siguen caminos diferentes.

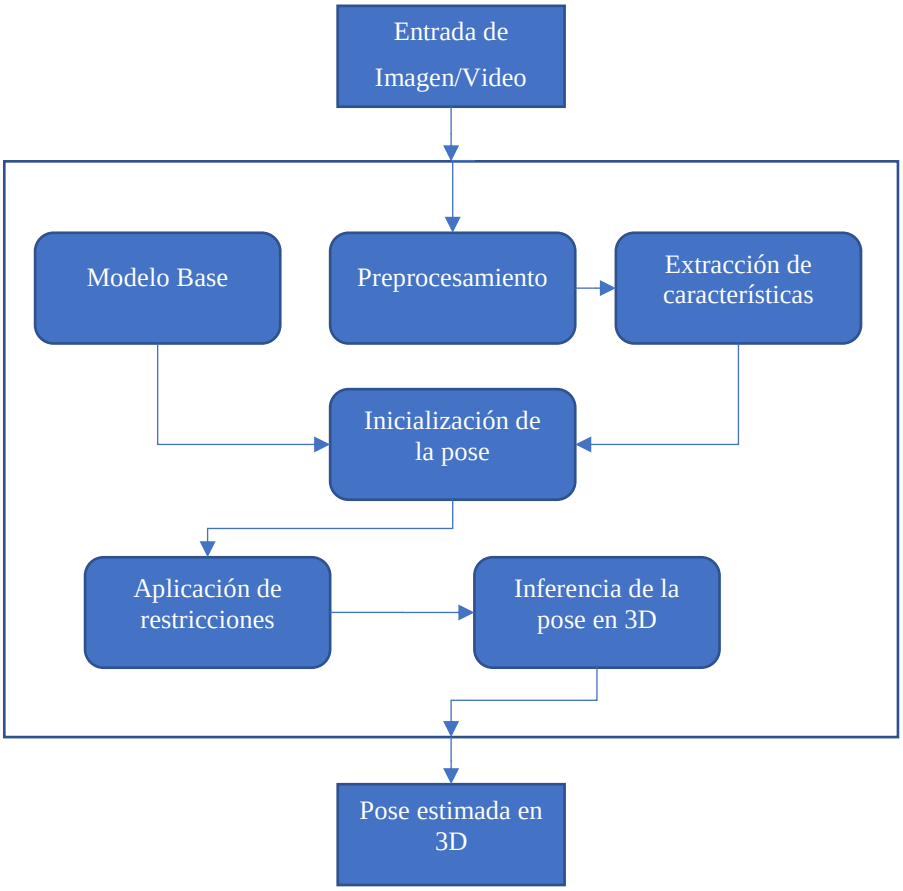

Fig. 2. ***Pasos comunes a los sistemas de estimación de la pose****. Dada una entra, la pose en 3D es obtenida mediante algunos o todos los pasos descritos (Fuente: Elaboración propia basada en la Fig.3 de Sarafianos et al.* (2016)).

Una manera de agrupar distintos enfoques es mediante el tipo de datos de entrada que utilizan. Para obtener una estimación en 3D se puede partir de imágenes RGB (), secuencias de imágenes RGB () o datos de profundidad (). Esta clasificación se puede dividir a su vez según el numero de cámaras que han obtenido los datos: imágenes o secuencias monoculares (una cámara) o multivista (más de una cámara).

Otro método de agrupación se basa en cómo interpretan la estructura del cuerpo. Así pues, si se utiliza un modelo del cuerpo conocido a priori, estaremos hablando de métodos con un enfoque generativo. Estos métodos utilizan la información estimada de los algoritmos para ir deformando el modelo inicial (Sminchisescu, 2011).

Una subcategoría de los modelos generativos son los modelos por partes. Éstos representan el modelo como un conjunto de partes del cuerpo, unidas por restricciones impuestas a las articulaciones en la estructura del esqueleto. Este modelo, ha sido ampliamente utilizado para la estimación de la pose en 3D en los últimos años (Amin *et al.*, 2013; Burenius, Sullivan and Carlsson, 2013).

Por su parte, los enfoques discriminativos son aquellos que no usan un modelo conocido a priori, ya que aprenden a mapear la relación entre las imágenes observadas y las poses humanas en 3D. Se pueden clasificar bajo dos tipos: los basados en el aprendizaje, que usan algoritmos para relacionar las imágenes observadas con el espacio de poses (Sedai, Bennamoun and Huynh, 2010); o los basados en el ejemplo, los cuales interpolan una postura a partir de varios candidatos almacenados que se obtienen mediante una búsqueda de similitud (Grauman, Shakhnarovich and Darrell, 2003).

La ventaja de los métodos generativos frente a los discriminativos es que, al generalizar bien, pueden inferir poses con mayor precisión y pueden manejar poses complejas. En cambio, los métodos discriminativos ofrecen mucha más rapidez y menos complejidad de cálculo.

Por último, podemos encontrar métodos híbridos, en cuyo caso combinan métodos generativos con discriminativos. La función de probabilidad obtenida de los métodos generativos se usa para verificar las hipótesis obtenidas de la función de mapeado de los métodos discriminativos (Rosales and Sclaroff, 2006).

1.3.2. Representaciones del cuerpo humano

El cuerpo humano está constituido por un sistema muy complejo de extremidades y articulaciones y representa un verdadero reto representar de manera realista las posiciones de dichas articulaciones en el espacio tridimensional. A pesar de esta dificultad, las técnicas de estimación automáticas proporcionan varios modelos que se pueden utilizar (*cf.* Fig. 3).

La representación más común de la estructura del cuerpo humano en tres dimensiones es la que se realiza mediante un esqueleto o figura de palos (*Pictorial Structures Model*, PSM por sus siglas en inglés). Estas figuras están compuestas por cilindros (partes del cuerpo) y puntos (articulaciones). A estos elementos se les aplican una serie de restricciones, como las restricciones anatómicas respecto al tamaño, a la simetría, a las proporciones o a las jerarquías en las articulaciones. Se emplean especialmente cuando se quiere representar solamente la pose y las relaciones entre sus partes (Sigal *et al.*, 2012).

Si lo que se desea es representar el cuerpo de manera más real en una pose, se utilizan los modelos de forma completa, los cuales muestran todo el contorno de la figura humana, tratada como una sola malla en 3D. Un ejemplo de este tipo lo constituyen los modelos SCAPE (Anguelov *et al.*, 2005).

Por último, existen modelos que combinan el realismo de los modelos SCAPE con las funcionalidades de los PSM, como el *Stitched Puppet* (Zuffi and Black, 2015). Cada parte del cuerpo tiene su propia forma, atributos y restricciones, los cuales se unen formando una malla completa y realista.

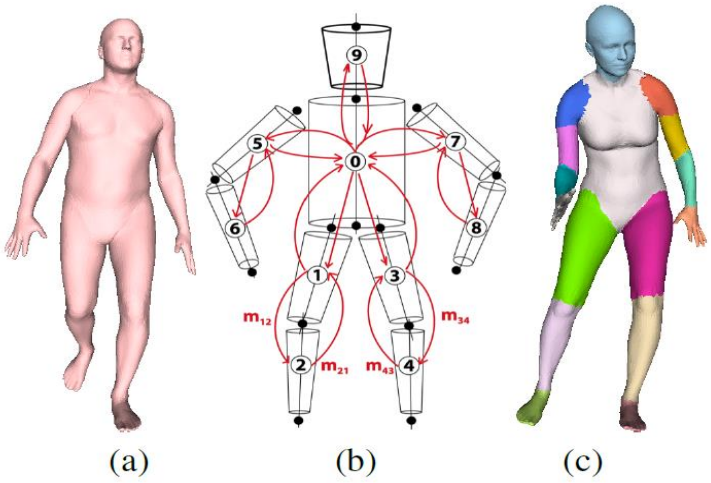

Fig. 3. ***Modelos del cuerpo en 3D****. (a) Representación realista de un modelo SCAPE. (b) Representación de un cuerpo mediante PSM. (c) Representación del modelo Stitched Puppet* (*Fuente: Zuffi y Black* (2015)).

1.3.3. Conjuntos de datos

Una limitación de la investigación existente sigue siendo la comparación de diferentes enfoques sobre conjuntos de datos comunes, así como la evaluación del desempeño de la exactitud. Este problema es abordado en *HumanEva I & II* (Sigal, Balan and Black, 2010), un conjunto de datos que contiene múltiples sujetos realizando acciones predefinidas.

Dicho conjunto contiene videos sincronizados desde distintas cámaras, así como valores 3D etiquetados, medidas cuantificables y un algoritmo base de seguimiento (*tracking*). Asimismo, está dividido en dos partes (I y II), cada una de las cuales contiene diferente número y tipo de cámaras, de movimientos, de tipos de datos y de sincronización. Además, está compuesto por subconjuntos de entrenamiento, validación y test. Cabe finalmente destacar que se trata del conjunto más usado en este ámbito de investigación, puesto que permite la comparación entre distintos métodos usando los mismos datos y unidades de error.

Existen otros conjuntos de datos significativos disponibles, como el *CMU Graphics Lab Motion Capture* (Carnegie Mellon University Graphics Lab, no date) o el *Human 3.6M* (Ionescu *et al.*, 2014).

Dichos conjuntos presentan limitaciones que todavía requieren ser solventadas. Actualmente, es impracticable proporcionar videos en entornos sin restricciones, debido a la dificultad de etiquetar de forma precisa y manual los datos 3D de los cuerpos que aparecen, sin limitar los escenarios o las acciones realizadas. Asimismo, resulta necesario recabar datos procedentes de personas con dimensiones antropomórficas distintas y, si es posible, con vestimentas variadas. Finalmente, también cabe recopilar información sobre escenas en las que aparezcan más de una persona, ya sea interaccionando entre sí o con objetos.

Estas limitaciones tratan de superarlas Chen *et al.* (2016) generando un conjunto de datos sintéticos. Este conjunto está formado por imágenes obtenidas con un algoritmo que mezcla modelos en 3D, poses humanas, imágenes de ropa y fondos, obteniendo una gran variedad de modelos antropomórficos diferentes, con ropa variada y poses distintas. La gran ventaja de generar datos de esta manera es que resulta mucho más fácil etiquetar los datos, ya que no hay que hacerlo manualmente.

En lo que concierne a los conjuntos de datos de imágenes de profundidad, desde el lanzamiento de la cámara Kinect, disponemos de una herramienta eficaz y barata para obtener la cantidad de datos necesarios. Debido a sus características, las escenas que se obtienen pertenecen a escenarios de interior. Un ejemplo de ello lo encontramos en el trabajo de Li, Zhang y Liu (2010) con su conjunto de datos *MSR Action 3D*, el cual contiene 20 acciones orientadas a los videojuegos y llevadas a cabo por siete personas distintas.

Con relación a los conjuntos de datos multicámara podemos encontrar *ReadingAct*, realizada por el departamento de computación de la Universidad de Reading (http://www.cvg.reading.ac.uk/). Éste contiene 19 tipos de actividades diarias realizadas en interior por 20 personas distintas y grabadas con la ayuda de dos cámaras de manera simultánea, una colocada de frente y otra de perfil al actor.

1.3.4. Métricas de evaluación

La gran variedad de retos que existen a la hora de estimar la pose humana ha llevado a que los distintos investigadores desarrollen diferentes métricas de evaluación, lo que ha dificultado la comparación entre distintos métodos.

Los autores del conjunto *HumanEva* (Sigal, Balan and Black, 2010) introdujeron el Error 3D (ε), que representa la distancia cuadrática media en 3D (medida en milímetros) entre el conjunto de marcadores virtuales correspondientes a los centros de las articulaciones estimadas y las reales:

$$\varepsilon(x, \hat{x}) = \frac{1}{M}\sum_{i=1}^{M}\|m_i(x) - m_i(\hat{x})\|$$

Donde x representa la pose real, $\hat{x}$ la pose estimada, M el numero de marcadores virtuales y $m_i(x)$ representa la posición en 3D del marcador virtual o real en la posición i.

Una métrica reciente es el Porcentaje de Partes estimadas Correctamente (*Percentage of Correctly estimated Parts*, PCP por sus siglas en inglés) aplicado al 3D (Belagiannis *et al.*, 2013). Se considera que una parte está correctamente estimada si:

$$\frac{\|S_n - \hat{S}_n\| + \|e_n - \hat{e}_n\|}{2} \leq \alpha\left\|S_n - \hat{S}_n\right\|$$

Donde S_n y e_n representan las coordenadas en 3D reales del punto inicial y final de la parte n, $\hat{S}_n$ y $\hat{e}_n$ sus respectivas estimaciones y α el parámetro que controla el umbral.

Otras métricas frecuentes obtienen el error en grados, como el Error Medio del Ángulo de la Articulación (*Mean Joint Angle Error*, MJAE por sus siglas en inglés) (Ning *et al.*, 2008). Este error mide la desviación media en grados de las articulaciones estimadas, frente a las etiquetadas en el conjunto de datos de la siguiente manera:

$$D(y, y') = \frac{1}{M}\sum_{i=1}^{M}\left|\left(y_i - y'_i\right) mod \pm 180°\right|$$

Donde y_i e y'_i son los vectores de pose de la articulación estimada y de la real, respectivamente, y M es el número de articulaciones. El resultado en grados se obtiene en el rango $[-180°, +180°]$.

2. Métodos de estimación a partir de una imagen 2D

La obtención de una configuración de puntos 3D a partir de una imagen RGB presenta tres grandes problemas que afectan a su rendimiento:

1. Algunas proyecciones similares en 2D pueden derivar de diferentes poses en 3D.
2. Algunos pequeños errores en la localización de características en 2D pueden tener un gran impacto en el espacio 3D.
3. La estimación se ve afectada por el problema de la alta dimensionalidad.

2.1. Estimación a partir de una imagen monocular

Obtener una pose 3D a partir de imágenes monoculares en 2D constituye una tarea especialmente difícil por diversos motivos, entre los que destacamos la alta no linealidad del movimiento humano, la gran variedad de apariencias y poses, los fondos aglomerados, las oclusiones y las proyecciones ambiguas. Los métodos explicados en este apartado estiman una pose en 3D explícitamente desde una sola imagen.

2.1.1. Métodos de deep-learning

Los métodos de *deep-learning* ya se empleaban con anterioridad para la extracción de la pose en 2D (Chen and Yuille, 2014), aunque también se han empezado a usar para la extracción en 3D.

Un ejemplo lo encontramos en el trabajo de Li y Chan (2014), quienes entrenan una red convolucional profunda siguiendo dos estrategias: (1) entrenar conjuntamente la regresión de la pose y los detectores de las partes del cuerpo y (2) entrenar la regresión de la pose usando una red previamente entrenada para detectar las partes del cuerpo. En su estudio

concluyen que la estimación de la pose es un problema de predicción estructurada, ya que, aunque no añaden restricciones a la red sobre las partes del cuerpo, ésta aprende sola las dependencias y relaciones entre las distintas partes del cuerpo.

Li, Zhang y Chan (2015) proponen una red neuronal profunda que recibe una imagen y una pose en 3D y devuelve un valor representando si la pose corresponde a la imagen. Para lograrlo utilizan una red neuronal con el fin de extraer características de la imagen, seguida de dos subredes que transforman estas características y la pose en una estructura de articulaciones embebidas, para finalmente obtener una puntuación sobre la imagen.

2.1.2. Detectores 2D

Para poder superar la dificultad y el coste de adquirir imágenes de poses con sus respectivas anotaciones en 3D, Yasin *et al.* (2016) proponen un método donde se utilizan dos orígenes de datos independientes, procedentes de imágenes con anotaciones de la pose en 2D y datos 3D precisos obtenidos a partir de capturas de movimiento. Dichos datos se usan durante el entrenamiento para proyectarlos en 2D mediante un modelo de regresión. Finalmente, para obtener una pose en 3D desde una imagen, primero se estima la pose en 2D, después se obtiene la pose 3D más probable que proyectará esta pose en 2D y, por último, se le minimiza el error para obtener la pose 3D final.

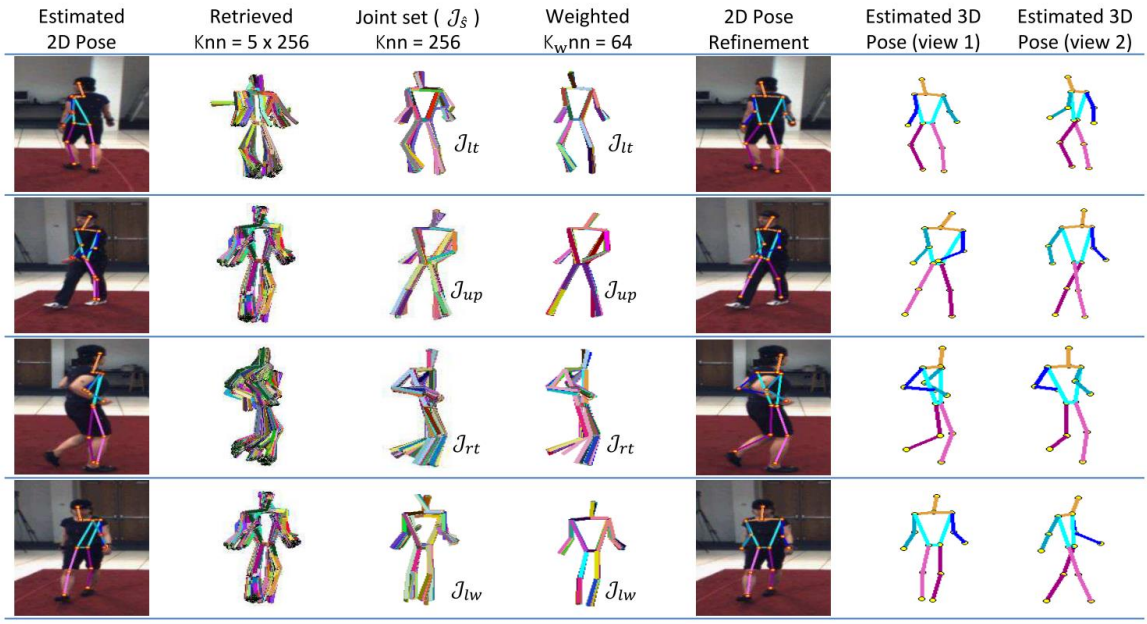

Fig. 4. ***Ejemplos del método propuesto por Yasin et al.* (2016).** *Se observan las estimaciones de la pose en varias iteraciones y el resultado final para cada imagen (Fuente: Yasin et al.* (2016)).

2.1.3. Estimación de los parámetros de la cámara

Para resolver la ambigüedad que surge al realizar una estimación a partir de una única imagen, algunos métodos estiman también la pose relativa de la cámara (Wang *et al.*, 2014). Estos métodos necesitan la posición de las articulaciones en 2D con el fin de estimar la pose en 3D y los parámetros de la cámara, los cuales se utilizan para eliminar la ambigüedad en las posibles poses estimadas.

2.1.4. Aproximaciones discriminativas

Un ejemplo de estos métodos lo encontramos en el trabajo de Kostrikov y Gall (2014), quienes proponen un barrido discriminativo profundo de un árbol de regresión. Tras extraer las características desde imágenes 2D de distinta profundidad, el método propuesto barre con un plano el volumen 3D de posiciones potenciales de las articulaciones y utiliza un árbol de regresión que aprende a mapear de 2D a 2D o de 3D a 3D las posiciones relativas de las articulaciones. Por ello, el método predice la posición 3D relativa de una articulación, dada la hipotética profundidad de la característica. Por último, utiliza un modelo con restricciones para afinar la estimación obtenida.

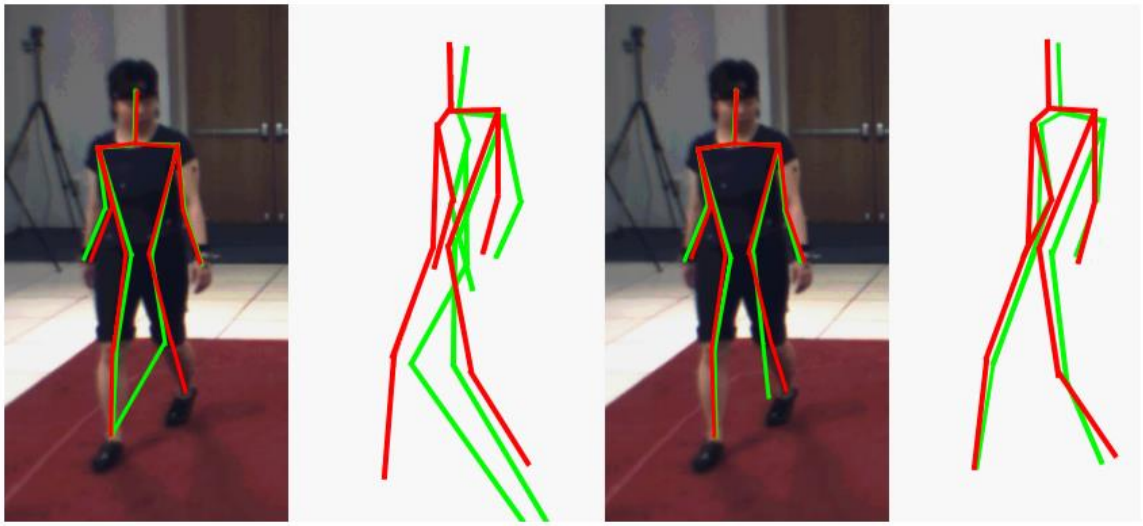

Fig. 5. ***Ejemplo del método propuesto por Kostrikov y Gall* (2014).** *Las dos imágenes de la izquierda corresponden a la estimación obtenida por los arboles de regresión. Las dos imágenes de la derecha corresponden a la estimación ajustada usando el modelo. En verde la pose estimada y en rojo los datos reales* (*Fuente: Kostrikov y Gall* (2014)).

2.2. Estimación a partir de una imagen en un escenario multicámara

En vez de usar un modelo en 3D, Amin *et al.* (2013) utilizan una proyección en 2D en cada vista utilizando estructuras pictóricas, a las cuales se les introduce restricciones de

correspondencia (apariencia y espacial) para tomar ventaja de las diferentes vistas y obtener así, mediante triangulación, la pose en 3D.

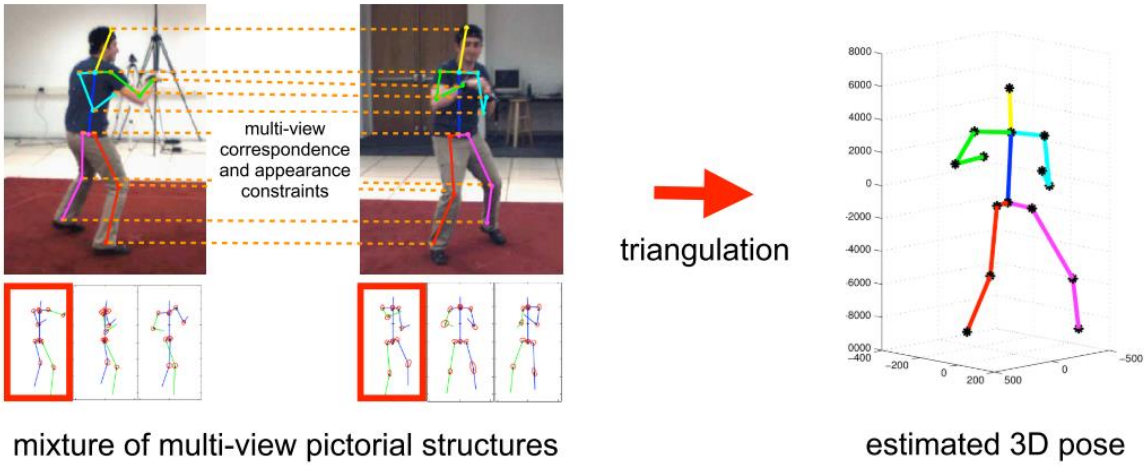

Fig. 6. ***Ejemplo del método propuesto por Amin et al.* (2013).** *A la izquierda, las estimaciones individuales de cada vista obtenidas de la mezcla de varios PSM. A la derecha, la estimación final obtenida por triangulación* (*Fuente: Amin et al.* (2013)).

3. Métodos de estimación a partir de una secuencia de imágenes 2D

Al tratar de estimar la pose en 3D a partir de una secuencia de imágenes se presentan algunas dificultades. Por ejemplo, en ambientes no controlados, podríamos apreciar grandes variaciones a lo largo de la secuencia, debido a cambios en el fondo de la imagen o en la iluminación. Incluso si nos encontráramos en un escenario controlado, la apariencia del cuerpo podría cambiar, por el movimiento de la ropa o porque se producen pequeñas variaciones en la posición o en la rotación de algunas partes del cuerpo, especialmente en las extremidades, debido a que el cuerpo no es una entidad rígida sino un conjunto de partes móviles.

3.1. Estimación a partir de una secuencia de imágenes monoculares

3.1.1. Métodos discriminativos

En el trabajo de Tekin *et al.* (2015), se utiliza la información espaciotemporal para reducir la ambigüedad en la profundidad. Dichos autores emplean dos redes convolucionales, primero para alinear los delimitadores del cuerpo a lo largo de las imágenes y, posteriormente, para ajustarlas con el fin de crear datos de volumen. A partir de ello, se obtienen descriptores HoG en 3D, para finalmente reconstruir la pose tridimensional

utilizando KRR (*Kernel Ridge Regression*) y KDE (*Kernel Dependency Estimation*). Mediante este método demuestran que cuando se tiene en cuenta información de múltiples imágenes consecutivas, las poses con auto-oclusión pueden ser estimadas con mayor precisión.

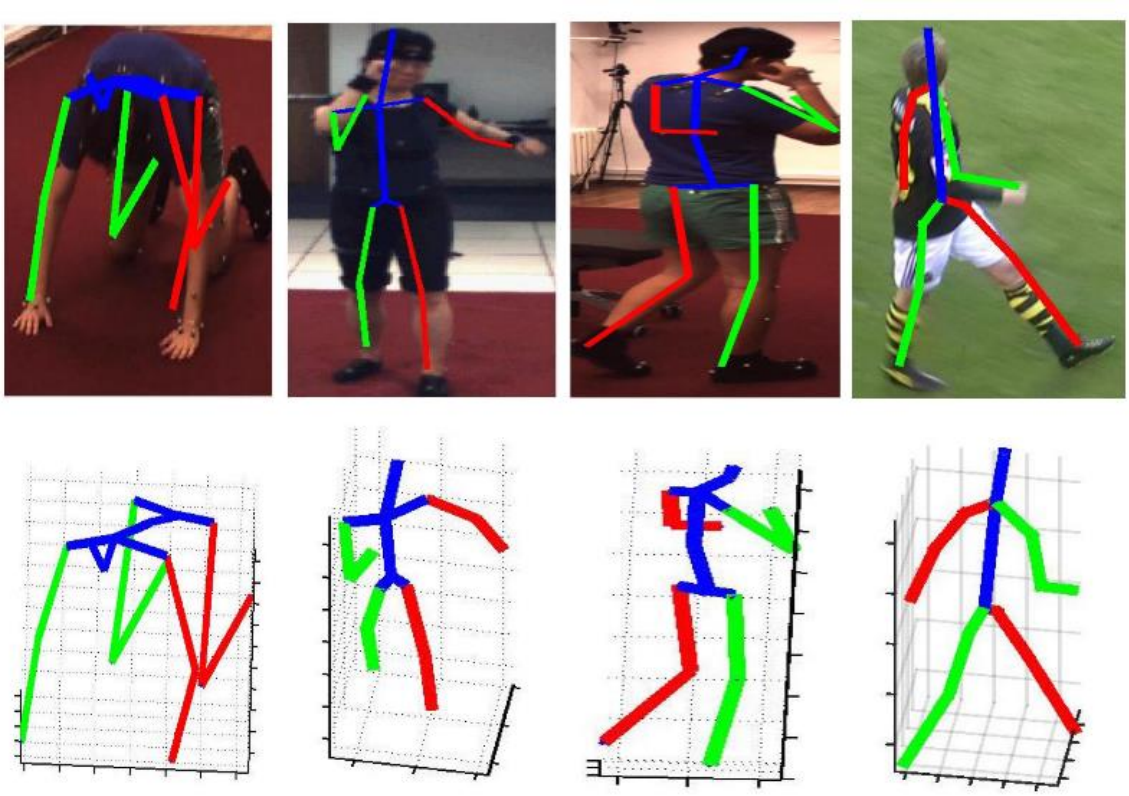

Fig. 7. ***Ejemplo del método propuesto por Tekin et al.*** **(2015).** *Poses estimadas correspondientes a imágenes de distintos conjuntos de datos (Fuente: Tekin et al.* (2015)).

3.1.2. Modelos de variables latentes

En este campo de investigación se utilizan a menudo variables latentes, puesto que a veces resulta difícil obtener estimaciones precisas a causa de las posibles oclusiones. Como proponen Tian *et al.* (2013), hacer uso de ellas contrarresta el sobreentramiento (*overfitting*) y la generalización deficiente.

Andriluka *et al.* (2010) desarrollan un método para estimar la pose 3D en escenas del mundo real, donde aparecen múltiples personas y, en ocasiones, tienen lugar oclusiones parciales o completas. El método hibrido que proponen consta de tres pasos. En primer lugar, a partir de un detector de partes en 2D, obtienen la localización de las articulaciones en las imágenes. En la segunda fase, mejoran la robustez de la estimación en 2D y habilitan la asociación temprana de datos mediante un *tracking* por detección en 2D. Finalmente, se

estima la pose en 3D mediante un proceso gaussiano jerárquico de variables latentes (hierarchical *Gaussian Process Latent Variable Model*, hGPLVM por sus siglas en inglés), el cual se combina con un modelo oculto de Markov (*Hidden Markov Model*, HMM por sus siglas en inglés).

Fig. 8. ***Ejemplos del método propuesto por Andriluka et al.*** **(2010).** *Tracking 3D de una escena en exterior* (*Fuente: Andriluka et al.* (2010)).

3.1.3. Algoritmo de filtrado de partículas

Sedai, Bennamoun y Huynh (2013) introducen un método hibrido que utiliza un filtro de partículas gaussiano. Dichos investigadores emplean un regresor de un proceso gaussiano que recibe un descriptor de silueta y produce varias poses. En la parte de *tracking*, el conjunto de poses obtenidas se mezcla con el filtro de partículas en cada imagen de la secuencia. En su trabajo muestran que su método no requiere de inicialización y que no pierde la pose durante el video.

3.1.4. *Tracking*

Los métodos que se engloban bajo esta categoría utilizan información temporal para seguir el movimiento del cuerpo a lo largo de la secuencia. Uno de estos métodos (Rius et

al., 2009) hace uso de un modelo dinámico específico para una acción, lo cual descarta configuraciones de la pose que no se adapten a esa acción en concreto. Después, dada la posición 2D de un conjunto de articulaciones, el modelo se combina con un *framework* de filtrado de partículas para estimar la pose.

Iqbal, Milan y Gall (2017) proponen en un método reciente, un algoritmo de seguimiento capaz de estimar la pose de varias personas. Para ello, representan las posiciones de las articulaciones detectadas en la secuencia mediante un grafo espaciotemporal y resuelven un problema de programación lineal en enteros para particionar el grafo en subconjuntos que corresponden a la trayectoria de cada persona.

Fig. 9. ***Ejemplos del método propuesto por Iqbal, Milan y Gall* (2017).** *Estimación de la pose para cada persona en el video. Cada color corresponde a la identidad de una persona* (*Fuente: Iqbal, Milan y Gall* (2017)).

3.2. Estimación a partir de una secuencia de imágenes en un escenario multicámara

El enfoque de Belagiannis *et al.* (2013) utiliza modelos 3D del cuerpo humano para estimar la pose de varios cuerpos en una secuencia de imágenes con múltiples cámaras. En su primer trabajo se encontraron dos problemas: por una parte, con la gran dificultad de identificar a cada persona; por otra parte, con la necesidad de solucionar las oclusiones que cada cuerpo provocaba a los de alrededor. Otro problema que abordaron era el espacio de estado complejo de alta dimensionalidad. Para ello, en vez de discretizarlo, utilizaron algoritmos de triangulación sobre las articulaciones del cuerpo correspondientes, obtenidas a partir de un detector de partes en 2D para cada par de vistas. El modelo que usaron fue un modelo pictórico en 3D (3DPSM), el cual infiere la pose articulada de varias personas, al mismo tiempo que resuelve la ambigüedad que surge por utilizar múltiples cámaras. Dicho modelo está basado en un campo condicional aleatorio y fuerza restricciones cinemáticas, de rotación y de colisión. Por último, la inferencia del 3DPSM se consigue mediante un algoritmo de propagación de creencias en bucle. En un trabajo posterior (Belagiannis *et al.*, 2015), dotaron al modelo 3DPSM de consistencia temporal. En él, primero identifican a cada

individuo mediante un algoritmo de seguimiento, que deriva en un espacio menor, y facilitan la inferencia mediante la penalización de los candidatos que difieran significativamente en cuanto a la geometría corporal.

Por su parte, Elhayek *et al.* (2015) proponen un método novedoso que permite hacer un seguimiento de la pose, tanto en interior como en exterior, usando tan solo dos o tres cámaras. Para cada articulación se parte de una red neuronal convolucional distinta en vistas a estimar los potenciales unarios creando un método discriminatorio de partes del cuerpo. Probabilísticamente se extraen restricciones mediante el uso de los potenciales y muestras ponderadas.

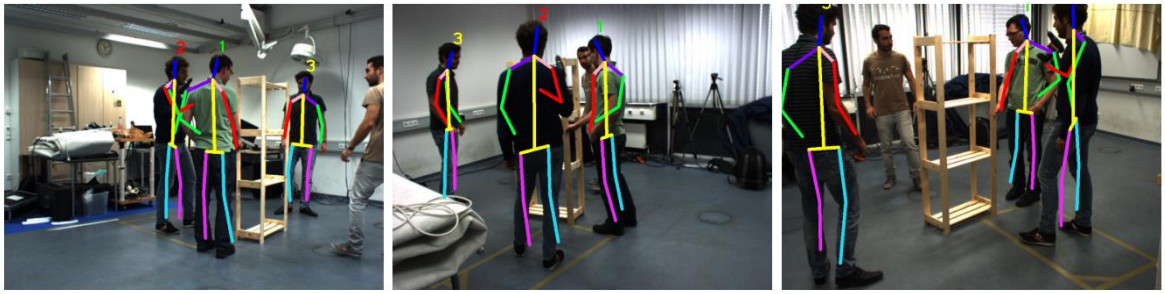

Fig. 10. ***Ejemplo del método propuesto por Belagiannis et al.* (2015).** *Pose estimada vista desde distintas cámaras* (*Fuente: Beleagiannis et al.* (2015)).

4. Métodos de estimación a partir de datos de profundidad

Partiendo de la base de que el cuerpo real es un objeto tridimensional, el hecho de utilizar información de profundidad (voxels) puede evitar el uso repetitivo de técnicas de proyección de los modelos en 3D a 2D, para comparar con las características extraídas de las imágenes bidimensionales. Una de las ventajas de usar la información en voxels es que permite el diseño de algoritmos más simples que pueden usar el conocimiento previo de las formas y tamaños de las partes del cuerpo humano. Por supuesto, esto revierte en un aumento del coste computacional, aunque ya se han desarrollado técnicas para mejorar la eficiencia (Cheung *et al.*, 2000).

4.1. Sensores de profundidad

Al hablar de datos de profundidad resulta necesario hacer mención a los tipos de cámaras a partir de las cuales éstos se pueden obtener. En este sentido, distinguimos:

1. Cámaras estéreo: infieren la estructura en 3D de una escena utilizando dos o más cámaras para emular la visión estéreo humana. Debido a la complejidad de los

cálculos geométricos requeridos, este tipo de cámaras son muy lentas y resultan no aptas para el uso de aplicaciones en tiempo real. Además, al tratarse de cámaras RGB, son sensibles a los cambios en la iluminación.

2. Cámaras TdV: las cámaras de *tiempo-de-vuelo* (TdV) estiman, desde una sola cámara, la distancia a la superficie de un objeto mediante un pulso de luz activo. Dichas cámaras calculan la profundidad midiendo el tiempo que tarda la luz en reflejar en el objeto. Aunque suelen tener poca resolución, los cálculos necesarios se obtienen rápidamente, pudiendo alcanzar un alto ratio de imágenes por segundo.
3. Sensores de luz estructurada: la cámara más representativa de este tipo es la Kinect v1. Está compuesta por un emisor de infrarrojos que emite un patrón de luz irregular y conocido, y un sensor que recoge el patrón deformado por los objetos de la escena. La información de profundidad se obtiene mediante triangulación. La principal ventaja de estas cámaras frente a las TdV es que son mucho más baratas, aunque obtienen imágenes de profundidad con huecos, ya que hay zonas de la escena que no pueden ser vistas por el emisor y el sensor al mismo tiempo.

4.2. Preprocesamiento

4.2.1. Substracción del fondo y de ruido

Una de las ventajas de las imágenes de profundidad es que resulta mucho más fácil separar los objetos del primer plano del fondo de la imagen. Un ejemplo de ello lo podemos encontrar en el trabajo de Schwarz *et al.* (2012), los cuales emplean una imagen estática del fondo tomada con anterioridad, para extraerla de la imagen en la que se encuentra la persona.

Por el contrario, uno de los inconvenientes es que pueden generar ruido, por lo tanto requieren de la aplicación de alguna técnica de reducción de ruido, como el filtrado de medias o las operaciones morfológicas (Wu, Zhu and Shao, 2012).

4.2.2. Llenado de huecos

Como se menciona anteriormente, las cámaras con visión estéreo (cámaras estéreo y cámaras de luz estructurada) proporcionan imágenes con huecos, es decir, lugares donde la información de profundidad es desconocida. Liu *et al.* (2016) proporcionan un método para paliar estos efectos mediante una combinación de filtros morfológicos y filtros de bloque cero.

4.3. Métodos generativos

Como se comenta en la introducción, los métodos generativos utilizan modelos del cuerpo parametrizados que se encajan en los datos de profundidad mediante esquemas de optimización.

Una primera aproximación presentada por Pekelny y Gotsman (2008) realizaba un seguimiento de los huesos individualmente utilizando ICP (*Iterative Closest Point*) sobre objetos rígidos articulados. Otra aproximación (Knoop, Vacek and Dillmann, 2009) basada en ICP genera correspondencias entre puntos 3D y 2D. En dicha propuesta, se elige un punto bidimensional mediante un detector de características y se define un rayo con todos los puntos tridimensionales que pueden ser proyectados sobre ese punto en 2D. El punto del modelo 3D más cercano a ese rayo es el que se emplea para generar una restricción tradicional. Aunque los autores de este método consiguen un rendimiento de 25 fps, solo se considera válida para posiciones sin oclusión.

Un método más reciente lo encontramos en el trabajo de Oyama *et al.* (2017), el cual plantea obtener la forma del cuerpo en 3D a partir de una sola imagen de profundidad. Para ello, utilizan un enfoque en dos partes. La primera parte consiste en encajar, de manera general, un modelo de una base de datos que concuerde con la pose de la imagen observada mediante deformación del esqueleto. Posteriormente, se encaja el modelo de manera más precisa mediante un editor de superficies de Laplace. En la segunda parte se usa un *Stitched Pupped* para recuperar los detalles de la forma perdidos en el proceso. Como ventaja, cabe destacar que este método puede obtener la forma del cuerpo, aunque lleve ropa y los datos contengan ruido.

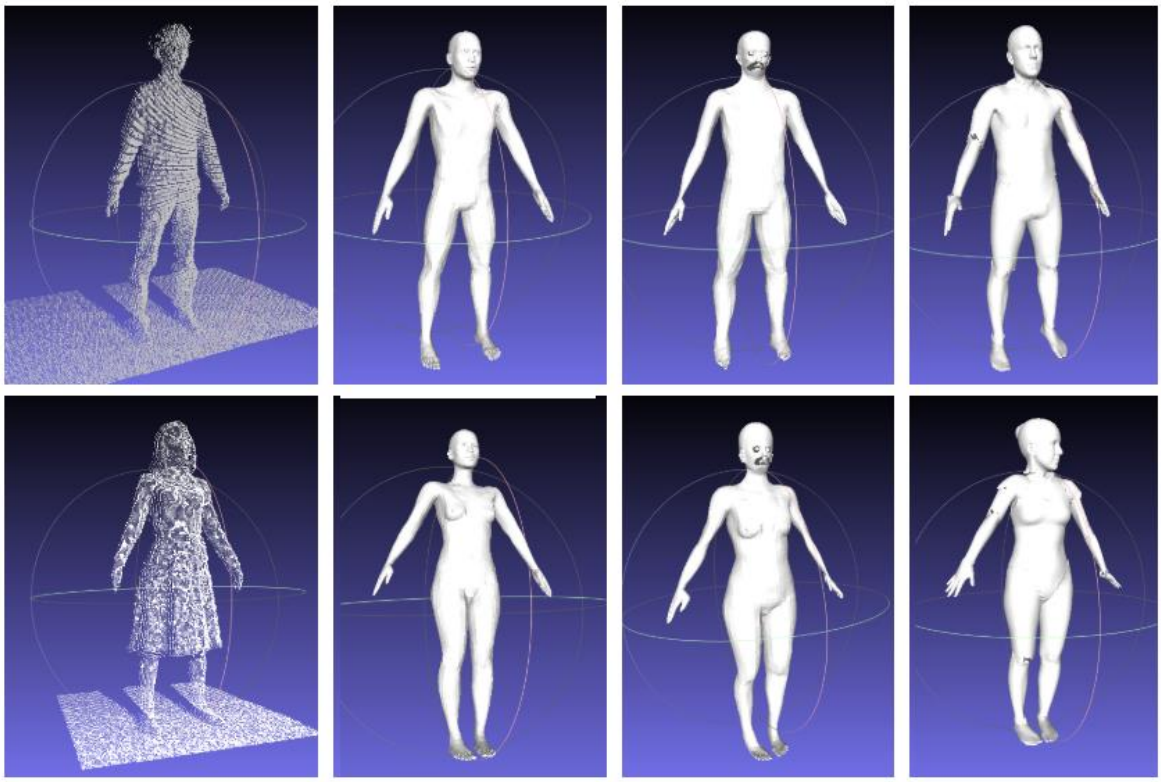

Fig. 11. ***Ejemplos del método propuesto por Oyama et al.*** **(2017).** *La primera imagen de cada fila corresponde a los datos de entrada, la segunda y tercera el estado de la estimación en varios puntos intermedios, y en último lugar la estimación final* (*Fuente: Oyama et al.* (2017)).

4.4. Métodos discriminativos

Estos métodos generalmente se centran en detectar ciertas características en los datos de profundidad para formar una hipótesis de la posición del cuerpo. Normalmente estos rasgos se aprenden a partir de un conjunto de datos de poses.

Plagemann *et al.* (2010) presentan un método que llaman extremos geodésicos. Se trata de puntos que representan los extremos del cuerpo (cabeza, manos y pies) calculados con un algoritmo iterativo de Dijkstra sobre un grafo obtenido mapeando las conexiones de todos los voxels.

Usando arboles de regresión, Girshick *et al.* (2011) estiman la posición de cada articulación dejando que cada voxel vote a una articulación. Tras descartar votos de voxels muy lejanos y de aplicar un estimador de densidad, pueden incluso estimar la posición probable de las articulaciones no visibles.

Uno de los métodos más eficientes a este respecto es el de Shotton *et al.* (2013). Su método puede predecir de manera rápida y precisa la posición en 3D de las articulaciones

del cuerpo a partir de una sola imagen de profundidad y sin usar información temporal. Dichos autores utilizan un algoritmo basado en reconocimiento de objetos e incluyen en el diseño una fase intermedia de reconocimiento de las partes de cuerpo, que mapea el difícil problema de la estimación de la pose en un problema más simple de clasificación por pixel. Gran parte de su éxito se debe al uso de una gran y extensa variedad de datos de entrenamiento, lo que permite al clasificador aprender a estimar las partes del cuerpo siendo invariable a la pose, la forma anatómica, la ropa, etc. El algoritmo genera propuestas confidenciales tridimensionales de varias articulaciones del cuerpo mediante la reproyección del resultado de la clasificación y la búsqueda de modos locales. De esta manera, pueden estimar correctamente a casi 200 fps la pose humana, siendo invariable en cuanto a la luminosidad y la escala, y pudiendo manejar algunos problemas de oclusión. Sin embargo, este método tiene una limitación: la distancia efectiva a la que trabaja es relativamente corta.

Alternativamente, proponen otro método que directamente aplica regresión para calcular la posición de las articulaciones (Shotton, Girshick, *et al.*, 2013). Comparándolo con su modelo anterior, este último consigue más precisión y resulta más limpio al no tener que realizar pasos intermedios. Sin embargo, se trata de un algoritmo más complejo y muchos hiperparámetros se tienen que optimizar contra un conjunto de validación.

Uno de los estudios más recientes (Park *et al.*, 2017) estima la pose del cuerpo humano a través de múltiples tipos de árboles de regresión aleatoria. Este método utiliza dichos árboles para realizar una votación sobre la localización de las articulaciones y, posteriormente, un árbol de verificación aleatoria para mejorar la precisión de los votos y determinar la posición final de las articulaciones. La particularidad de este estudio es que se centra en poses de jugadores de golf.

Otro método especializado en extraer poses en quirófanos (Kadkhodamohammadi *et al.*, 2017) utiliza una red convolucional como detector de partes, combinado con un modelo 3D de deformación por pares.

Fig. 12. ***Ejemplos del método propuesto por Kadkhodamohammadi et al.*** **(2017).** *Los esqueletos aceptados se muestran el naranja y los rechazados en morado (Fuente: Kadkhodamohammadi et al.* (2017)).

4.5. Métodos híbridos

Combinando las aproximaciones anteriores, los métodos híbridos intentan obtener la estabilidad y la coherencia temporal de los métodos generativos y la robustez de inferir incluso articulaciones ocluidas de los métodos discriminativos.

Un primer método presentado por Ganapathi *et al.* (2010), combina el método de extremos geodésicos con un esquema generativo optimizado basado en ICP. Otro método, el cual aplica la estimación de la pose al reconocimiento del movimiento, parte de extremos geodésicos (Krüger *et al.*, 2010). Utiliza las posiciones de la cabeza, manos y pies como índice en una base de datos de poses. Una vez obtenida la pose de la base de datos y la pose obtenida en el *frame* anterior del video, se les realiza una optimización local a cada una. A continuación, deciden, basándose en la distancia Hausdorff, qué pose describe mejor la imagen de profundidad observada.

Otros métodos híbridos utilizan un método discriminativo para inicializar un método de seguimiento generativo o para recuperar el seguimiento cuando falla en algún *frame* (Wei, Zhang and Chai, 2012).

Finalmente, en el trabajo realizado por He *et al.* (2015) se propone un nuevo extractor de características (3DLSC) a partir de la silueta del cuerpo. Dichos autores incorporan un

modelo gráfico del cuerpo en un árbol de regresión junto con el extractor, con el fin de aprender estructuras del cuerpo y localizar articulaciones.

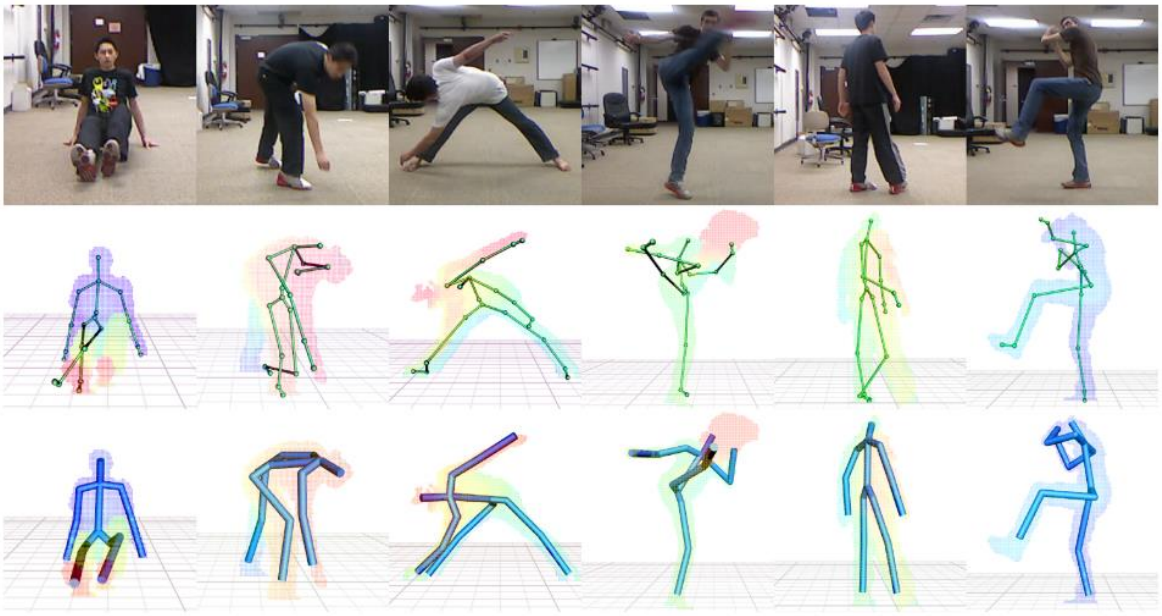

Fig. 13. ***Ejemplo del método propuesto por Wei, Zhang y Chain (2012).*** *Arriba las imágenes originales. En el centro la pose obtenida por la Kinect. Abajo la pose obtenida por su algoritmo (Fuente: Wei, Zhang y Chain (2012)).*

5. Análisis y comparativa

5.1. Problemas y retos

En este apartado procedemos a discutir los retos a los que se enfrentan los distintos métodos comentados anteriormente, así como su adecuación o no a ciertas aplicaciones.

5.1.1. Precisión de los modelos del cuerpo

Como se ha visto en apartados anteriores, muchos métodos dependen de trabajar con un modelo del cuerpo. Aunque existen modelos de distinta complejidad, ésta mayormente depende del objetivo de la aplicación. Mientras que algunos enfoques solo están interesados en obtener las posiciones de ciertos puntos del cuerpo, como las extremidades (Plagemann *et al.*, 2010) o las articulaciones (Wang *et al.*, 2014), otros tratan de extraer más información como los ángulos de las articulaciones (Ganapathi *et al.*, 2012) o incluso conseguir una superficie completa de la persona (Ye *et al.*, 2012).

En general se asume que para obtener modelos más complejos se tienen que crear a priori mediante modelado manual o equipamiento especial, como escáneres láser de cuerpo

completo. Aunque puede ser factible en ámbitos como la producción de películas o de videojuegos de grandes compañías, no lo es para otros escenarios con menos recursos. Los modelos SCAPE (Anguelov *et al.*, 2005) o *Stitched Pupped* (Zuffi and Black, 2015) proporcionan una solución a este problema, tanto si se trabaja con datos de entrada en 2D como en 3D.

Sin embargo, todavía existen problemas para algunas aplicaciones, como por ejemplo en las aplicaciones de realidad aumentada, donde se tenga que adaptar un objeto virtual al contorno de la persona (por ejemplo, probarse ropa de manera virtual antes de comprarla). Para ello se necesitaría un modelo preciso que, aunque se podría obtener (Weiss, Hirshberg and Black, 2011), no sería factible en una aplicación en tiempo real.

5.1.2. Ambigüedad rotacional

Un problema inherente de trabajar con datos de profundidad es que no contienen información suficiente por sí solos como para determinar la correcta orientación de objetos con simetría rotacional, como las extremidades del cuerpo. Aunque se puede obviar el problema si se trabaja con modelos simples del cuerpo, o sobre imágenes independientes, se debe tener en cuenta si se pretende obtener un modelo preciso o existen relaciones temporales entre distintas estimaciones a lo largo de una secuencia.

5.1.3. Oclusiones

Uno de los problemas más grandes y, por consiguiente, más abordado por los autores, son las oclusiones. Cuando una parte del cuerpo queda fuera del alcance del sensor porque tiene algo delante que lo oculta, se dice que la parte está ocluida. Esto puede deberse a que otras partes del mismo cuerpo interfieren (auto-oclusión) o a que hay objetos u otras personas en medio.

Una solución obvia a este problema es el uso de más de un sensor (Amin *et al.*, 2013; Burenius, Sullivan and Carlsson, 2013) para reducir la posibilidad de que alguna parte quede fuera del alcance de los sensores. Aunque efectivos, estos métodos generalmente sufrirán penalizaciones en cuando al rendimiento al tener que procesar más datos, privándolos en algunos casos de funcionar en tiempo real. También tendrán un mayor impacto económico al tener que disponer de un mayor equipamiento físico.

Otra solución podría ser utilizar modelos del cuerpo más complejos, que dispongan de restricciones anatómicas y/o cinéticas, que permitan establecer el modelo sin tener la

posición de todas las articulaciones, como el *Stitched Puppet* (Zuffi and Black, 2015). Utilizando información adicional, el modelo sería capaz de reconstruir los datos restantes (Oyama *et al.*, 2017).

5.1.4. Ambigüedad de la proyección 3D

Un problema intrínseco de estimar una pose tridimensional a partir de datos en 2D es que ciertas poses bidimensionales muy parecidas pueden ser proyecciones de poses tridimensionales muy distintas. Muchos de los métodos de estimación a partir de imágenes RGB atenúan este problema utilizando distintas aproximaciones.

Una de ellas consiste en utilizar más de una cámara. Al disponer de más de una proyección, el número de poses en 3D que pueden generar esas proyecciones se reduce. Además, se pueden añadir restricciones a las posibles poses en 3D para descartar aquellas que sean, por ejemplo, anatómicamente imposibles (Amin *et al.*, 2013).

Otras soluciones utilizan redes neuronales o arboles de decisión para aprender a mapear correctamente entre poses bidimensionales y tridimensionales mediante grandes volúmenes de datos desde los cuales entrenar (Kostrikov and Gall, 2014; Li and Chan, 2014)

5.1.5. Estimación de la pose en varias personas

Si se pretende estimar la pose de más de una persona al mismo tiempo, se debe tener en consideración si se quiere asociar cada pose a su correspondiente cuerpo en la imagen. Este problema se ve acentuado si se están utilizando secuencias de imágenes como datos de entrada. El algoritmo, por lo tanto, necesita identificar cada pose para mantener una cohesión temporal a lo largo de la estimación. Iqbal, Milan y Gall (2017) abordan este reto mediante un algoritmo de *tracking* con la ayuda de un grafo espaciotemporal.

5.2. Comparaciones

En primer lugar, se proporciona una tabla resumen de los métodos detallados en las secciones 2, 3 y 4 (*cf.* Anexo). Esta tabla resume los datos de entrada, si las imágenes son monoculares o multivista, si el método es generativo, discriminativo o híbrido, el tipo de molde corporal usado, los algoritmos principales o más relevantes del método y los conjuntos de datos utilizados.

A continuación, se presentan diversas tablas comparando los resultados de distintos métodos sobre los conjuntos de datos más utilizados. Cabe destacar que los datos han sido

extraídos de cada uno de los trabajos, según informan sus autores. En los casos que se proporcionaba más de un dato se ha obtenido una media aritmética de los diversos experimentos para poder representar el método en un único valor. En otros casos en cambio, los datos se proporcionaban en gráficos, haciendo difícil la extracción precisa de los valores, por lo que se ha obtenido un valor aproximado de manera visual. En algunos trabajos no se proporcionaban datos suficientes para poder establecer un valor comparativo.

En la Tabla 1 podemos observar aquellos métodos que reciben como entrada una imagen o secuencia de imágenes en 2D desde una sola cámara. Se ha elegido el Error 3D (ε) en mm como métrica al ser la más común en estos casos. Se ha realizado la comparación sobre los conjuntos *HumanEva* y *Human3.6M*, concretamente con los subconjuntos pertenecientes a la acción de caminar.

En la Tabla 2, siguiendo los mismos parámetros comparativos que la tabla anterior, encontramos los métodos que reciben como entrada una imagen o secuencia de imágenes en 2D desde más de una cámara.

Finalmente, en la Tabla 3 se exponen los métodos que reciben como entrada datos de profundidad. Se han incluido las métricas siguientes: el porcentaje de poses correctamente estimadas (PCP), por ser la más común, y el Error 3D (ε), para poder comparar con las tablas anteriores. No se ha dividido la tabla en conjuntos de datos como las anteriores debido a que cada método se basaba en un conjunto distinto, siendo en la gran mayoría, creado por los autores del trabajo.

Tabla 1. ***Métodos basados en imágenes o secuencias 2D monoculares.*** *Métrica: Error 3D (ε). Subconjunto de los datasets: Acción de caminar.* (*Fuente: Elaboración propia basada en los datos proporcionados por los autores*).

Año	***Método***	***Características***	***HumanEva***	***Human3.6M***
2017	Iqbal, Milan y Gall	*Tracking* + grafo espaciotemporal	---	---
2016	Yasin *et al*. (2016)	Imágenes 2D + capturas 3D	40.51 mm	76.03 mm
2015	Tekin *et al*. (2015)	Redes convolucionales + KRR y KDE	**37.3 mm**	**64.15 mm**
2014	Kostrikov y Gall (2014)	Arboles de regresión + PSM	75.86 mm	115.7 mm
	Li, Chan y Sijin (2014)	Red convolucional profunda	---	84.92 mm
	Wang *et al*. (2014)	Estimación y uso de parámetros de la cámara	77.6 mm	---
2013	Sedai *et al*. (2013)	Filtro gaussiano + *tracking*	53.1 mm	---
	Tian *et al*. (2013)	Variables latentes	---	---
2010	Andriluka *et al*. (2010)	Estimación 2D + *tracking* + hGPLVM y HMM	104 mm	---
2009	Rius *et al*. (2009)	Modelo específico para la acción	53.91 mm	---

Tabla 2. ***Métodos basados en imágenes o secuencias 2D multivista.*** *Métrica: Error 3D (ε). Subconjunto de los datasets: Acción de caminar.* (*Fuente: Elaboración propia basada en los datos proporcionados por los autores*).

Año	***Método***	***Características***	***HumanEva***	***Human3.6M***
2015	Belagiannis *et al*. (2015)	*Tracking* + 3DPSM	**53.1 mm**	---
	Elhayek *et al*. (2015)	Redes convolucionales + *tracking*	---	---
2013	Belagiannis *et al*. (2013)	Triangulación + 3DPSM	104 mm	---

Tabla 3. ***Métodos basados en imágenes o secuencias de profundidad.*** *Métricas: PCP y Error 3D (ε).* (*Fuente: Elaboración propia basada en los datos proporcionados por los autores*).

Año	***Método***	***Características***	***PCP***	***Error 3D***
2017	Oyama *et al*. (2017)	*Stitched Puppet* + base de datos	**---**	69.37 mm
	Park *et al*. (2017)	Árboles de regresión	**97.7%**	**31.04 mm**
2015	He *et al*. (2015)	Árboles de regresión + 3DLSC	97.5%	---
2013	Shotton *et al*. (2013)	Reconocimiento de partes	93%	---
	Shotton, Girshick *et al*. (2013)	Árboles de regresión	96%	---
2012	Wei, Zhang y Chai (2012)	*Tracking* + 3DPSM	95%	---
2011	Girshick *et al*. (2011)	Árboles de regresión	80%	---
2010	Ganapathi *et al*. (2010)	Extremos geodésicos	---	75mm
	Krüger *et al*. (2010)	Extremos geodésicos + base de datos	---	---
	Plagemann *et al*. (2010)	Extremos geodésicos	---	---
2009	Knoop, Vacek y Dillman (2009)	*Tracking* + ICP	---	---

6. Discusión

6.1. Problemas en la comparación

Uno de los objetivos del TFG era comparar entre los distintos métodos para ver cuál podría ser más adecuado para su aplicación en el proyecto de investigación. Sin embargo, existen ciertos problemas que evitan que estas comparaciones se puedan realizar correctamente. Como se ha visto en la introducción, existen distintos conjuntos de datos y métricas en los cuales medir la eficacia de un método de estimación de la pose. El problema principal viene dado porque los autores no se ponen de acuerdo en utilizar una metodología común, debido en muchos casos a que sus métodos tienen aplicaciones muy concretas y por

lo tanto necesitan de conjuntos de datos que generan ellos mismos (Kadkhodamohammadi *et al.*, 2017; Park *et al.*, 2017).

Observando las tablas anteriores podemos encontrar varios problemas que dificultan la tarea de poder comparar y decidir que métodos aportan mejores resultados. En las primeras tablas (*cf.* Tabla 1 y Tabla 2), el estar segmentadas para comparar sobre los conjuntos de datos más comunes impide comparar con métodos que no cumplan estos requisitos (Tian *et al.*, 2013; Iqbal, Milan and Gall, 2017). En la última tabla (*cf.* Tabla 3), nos encontramos dos problemas importantes. El primero es que la gran mayoría de métodos obtienen sus resultados de precisión de sus propios conjuntos de datos. El segundo problema reside en que la mayoría de los métodos no indican a partir de que error (ε) consideran una estimación como válida, por lo que un porcentaje de precisión no es suficiente para comparar si no se acompaña de un valor que pueda indicar que estimaciones se dan por buenas y cuáles no.

6.2. Elección de un método para el proyecto de investigación

Teniendo en cuenta todo lo anterior, la única solución para poder comparar y encontrar de manera precisa y correcta el método más adecuado para el proyecto de investigación, sería aplicar los métodos a un conjunto de datos especifico para nuestro problema y evaluar los resultados. Aunque no se puede proponer métodos que sepamos con seguridad que funcionen bien en nuestro escenario, si se puede indicar los candidatos más viables para realizar dicho estudio.

Recordemos pues, las características del reto que se plantea: estimación de la pose humana, al realizar una acción concreta, a partir de una secuencia de imágenes RGB-D obtenidas con la cámara Kinect.

Basándose en estas características, se proponen los siguientes candidatos para estudio:

1. **Wei, Zhang y Chai (2012).** Un buen punto de partida seria este método basado en *tracking*, ya que los autores evalúan su algoritmo enfrentándolo al de la Kinect (que es precisamente nuestro objetivo), obteniendo buenos datos de precisión y mejora respecto a este último. Además, parte de las mismas características que nuestro problema.
2. **He *et al*. (2015) y Park *et al.* (2017):** Siendo de los métodos más actuales y con los mejores resultados, resultan muy buenos candidatos para evaluar su funcionamiento,

aunque no tomen ventaja de la información temporal, sí hacen uso de los datos de profundidad.

3. **Tekin *et al.* (2015):** Con una precisión (en Error 3D) similar a los métodos anteriores, este método puede establecer la pose usando únicamente la información RGB. El método resulta muy sensible a los hiperparámetros, aumentando su complejidad a la hora de adaptarlo a nuestro problema.
4. **Rius *et al.* (2009):** Por último, este método, aunque con mayor error, permite especializarse en obtener la pose de una acción muy concreta. Al ser este nuestro caso, se podría implementar y comprobar si se obtienen buenos resultados.

6.3. Estado del arte

Existen varias cuestiones que no se han abordado al realizar la recopilación de métodos. La primera es que solo se han tenido en cuenta aquellos métodos que estiman la pose completa del cuerpo sin tener en cuenta la acción realizada. A pesar de eso, se han incluido trabajos que se centran en una acción o un escenario concreto, ya sea por su relevancia o sus buenos resultados.

Otra cuestión importante es la falta de artículos recientes incluidos en la recopilación. Está ligado con el problema anterior, y es debido a que los últimos estudios han dejado de centrarse en la estimación de la pose humana completa. Bien porque se han centrado en el análisis de partes más concretas del cuerpo, como las manos (Ji *et al.*, 2017; Chen *et al.*, 2018) o la cabeza (Veronese *et al.*, 2017), o bien porque han pasado a utilizar la estimación de la pose como una herramienta para el análisis del movimiento o las acciones (Jalal, Kamal and Kim, 2018), quedando fuera del alcance de este trabajo.

6.4. Trabajos futuros

Por lo tanto, como trabajos a realizar en el futuro se propone: la realización de un estudio sobre los resultados obtenidos al aplicar los métodos propuestos al problema que plantea el proyecto de investigación, y en segundo lugar, si es posible, la realización de una metodología que ayude a estandarizar el modo en que los autores evalúan los trabajos de estimación de la pose y permita una comparación real y mas general de sus métodos.

7. Conclusiones

Habiendo aportado una recopilación de los métodos realizados hasta la fecha, más destacados y significativos en el campo de la estimación de la pose humana en 3D, podemos concluir lo siguiente:

Después de analizar los problemas más importantes a la hora de estimar la pose (ambigüedades, oclusiones, etc.), se puede observar que muchos autores han conseguido de una manera u otra, desarrollar métodos capaces de superarlos (Amin *et al.*, 2013; Zuffi and Black, 2015; Oyama *et al.*, 2017), alcanzando en los últimos años algoritmos con muy buenos resultados (Park *et al.*, 2017). Debido a esto, trabajos más recientes no se están centrando en el reto de estimar la pose en general, sino que optan por un enfoque concreto en ciertas partes del cuerpo (Chen *et al.*, 2018), así como del estudio del movimiento y las acciones (Jalal, Kamal and Kim, 2018).

Una vez se han recopilado, analizado y comparado los distintos métodos, se ha llegado a la conclusión que no existe una metodología estándar entre los autores a la hora de evaluar sus trabajos, dificultando en gran medida la comparación y elección de métodos para su aplicación en un problema concreto. Esto es debido al uso de conjuntos de datos especializados y a la falta de información en cuanto a los resultados aportados por la mayoría de los autores.

Por último, aunque no se pueda asegurar el correcto funcionamiento de un método aplicado al problema que nos atañe, si que se puede proponer los métodos que mejor se adapten al reto que plantea el proyecto de investigación. Siendo así el trabajo de Wei, Zhang y Chai (2012) el mejor candidato, junto con los métodos de He *et al.* (2015) y Park *et al.* (2017).

8. Referencias

Amin, S. *et al.* (2013) 'Multi-view Pictorial Structures for 3D Human Pose Estimation', *Procedings of the British Machine Vision Conference 2013*, p. 45.1-45.11. doi: 10.5244/C.27.45.

Andriluka, M., Roth, S. and Schiele, B. (2010) '4. Monocular 3d pose estimation and tracking by detection', *Computer Vision and Pattern Recognition CVPR 2010*, (2), pp. 623–630. doi: 10.1109/CVPR.2010.5540156.

Anguelov, D. *et al.* (2005) 'SCAPE: Shape Completion and Animation of People', *SIGGRAPH '05 ACM SIGGRAPH 2005 Papers*, pp. 408–416. doi: 10.1007/978-3-642-37484-5_12.

Belagiannis, V. *et al.* (2013) '3D Pictorial Structures for Multiple Human Pose Estimation', (mm). doi: 10.1109/CVPR.2014.216.

Belagiannis, V. *et al.* (2015) 'Multiple human pose estimation with temporally consistent 3D pictorial structures', *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 8925, pp. 742–754. doi: 10.1007/978-3-319-16178-5_52.

Bhaltilak, K., Kaur, H. and Khosla, C. (2014) 'Human Motion Analysis with the Help of Video Surveillance: A Review', *International Journal of*, 4(9), pp. 245–249. Available at: http://search.proquest.com/openview/a15b3ebe69df860fb4b1f300e628c58b/1?pq-origsite=gscholar&cbl=2032130.

Burenius, M., Sullivan, J. and Carlsson, S. (2013) '3D pictorial structures for multiple view articulated pose estimation', *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 3618–3625. doi: 10.1109/CVPR.2013.464.

Carnegie Mellon University Graphics Lab (no date) *Motion Capture Database*, *Motion Capture Database*. Available at: http://mocap.cs.cmu.edu/.

Chen, T. Y. *et al.* (2018) 'Learning a deep network with spherical part model for 3D hand pose estimation', *Pattern Recognition*. Elsevier Ltd, 80, pp. 1–20. doi: 10.1016/j.patcog.2018.02.029.

Chen, W. *et al.* (2016) 'Synthesizing training images for boosting human 3D pose

estimation', *Proceedings - 2016 4th International Conference on 3D Vision, 3DV 2016*, pp. 479–488. doi: 10.1109/3DV.2016.58.

Chen, X. and Yuille, A. (2014) 'Articulated Pose Estimation by a Graphical Model with Image Dependent Pairwise Relations', pp. 1–9. Available at: http://arxiv.org/abs/1407.3399.

Cheung, G. K. M. *et al.* (2000) 'A real time system for robust 3D voxel reconstruction of human motions', *Proceedings IEEE Conference on Computer Vision and Pattern Recognition. CVPR 2000 (Cat. No.PR00662)*, 2(June 2014), pp. 714–720. doi: 10.1109/CVPR.2000.854944.

Condell, J., Moore, G. and Moore, J. (2006) 'Software and methods for motion capture and tracking in animation', *The 2006 International Conference on Computer Graphics and Virtual Reality*, p. 7 pp.

Elhayek, A. *et al.* (2015) 'Efficient ConvNet-based marker-less motion capture in general scenes with a low number of cameras', *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 07–12–June, pp. 3810–3818. doi: 10.1109/CVPR.2015.7299005.

Ganapathi, V. *et al.* (2010) 'Real time motion capture using a single time-of-flight camera', *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR2010)*, pp. 755–762. doi: 10.1109/CVPR.2010.5540141.

Ganapathi, V. *et al.* (2012) 'Real-Time Human Pose Tracking from Range Data', *Eccv (6)*, pp. 738–751.

Gavrila, D. . M. (1999) 'The Visual Analysis of Human Movement: A Survey', *Computer Vision and Image Understanding*, 73(1), pp. 82–98. doi: 10.1006/cviu.1998.0716.

Girshick, R. *et al.* (2011) 'A Survey of Efficient Regression of General-Activity Human Poses from Depth Images', pp. 415–422. Available at: http://arxiv.org/abs/1709.02246.

Grauman, Shakhnarovich and Darrell (2003) 'Inferring 3D structure with a statistical image-based shape model', *Proceedings Ninth IEEE International Conference on Computer Vision*, (Iccv), pp. 641–647 vol.1. doi: 10.1109/ICCV.2003.1238408.

He, L. *et al.* (2015) 'Depth-images-based pose estimation using regression forests and graphical models', *Neurocomputing*, 164, pp. 210–219. doi: 10.1016/j.neucom.2015.02.068.

Ionescu, C. *et al.* (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), pp. 1325–1339. doi: 10.1109/TPAMI.2013.248.

Iqbal, U., Milan, A. and Gall, J. (2017) ‘PoseTrack: Joint multi-person pose estimation and tracking’, *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, 2017–Janua, pp. 4654–4663. doi: 10.1109/CVPR.2017.495.

Jalal, A., Kamal, S. and Kim, D.-S. (2018) ‘Detecting complex 3D human motions with body model low-rank representation for real-time smart activity monitoring system’, *KSII Transactions on Internet and Information Systems*, 12(3), pp. 1189–1204. doi: 10.3837/tiis.2018.03.012.

Ji, Y. *et al.* (2017) ‘Hierarchical topology based hand pose estimation from a single depth image’, *Multimedia Tools and Applications*. Multimedia Tools and Applications, pp. 1–16. doi: 10.1007/s11042-017-4651-8.

Kadkhodamohammadi, A. *et al.* (2017) ‘A Multi-view RGB-D Approach for Human Pose Estimation in Operating Rooms’. Available at: http://arxiv.org/abs/1701.07372.

Knoop, S., Vacek, S. and Dillmann, R. (2009) ‘Fusion of 2d and 3d sensor data for articulated body tracking’, *Robotics and Autonomous Systems*. Elsevier B.V., 57(3), pp. 321–329. doi: 10.1016/j.robot.2008.10.017.

Kondoři, F. A. (2014) *Bring Your Body into Action*. Available at: http://umu.diva-portal.org/smash/get/diva2:716122/FULLTEXT01.pdf.

Kostrikov, I. and Gall, J. (2014) ‘Depth Sweep Regression Forests for Estimating 3D Human Pose from Images’, *Bmvc*, pp. 1–13.

Krüger, B. *et al.* (2010) ‘Fast Local and Global Similarity Searches in Large Motion Capture Databases’, *Eurographics / ACM SIGGRAPH Symposium on Computer Animation*, pp. 1–10. doi: 10.2312/SCA/SCA10/001-010.

Li, S. and Chan, A. B. (2014) ‘3D Human Pose Estimation from Monocular Images with Deep Convolutional Neural Network’, *Asian Conference on Computer Vision (ACCV)*, pp. 332–347. doi: 10.1007/978-3-319-16808-1_23.

Li, S., Zhang, W. and Chan, A. B. (2015) ‘Maximum-Margin Structured Learning with Deep

Networks for 3D Human Pose Estimation', *Iccv*, pp. 2848–2856. doi: 10.1109/ICCV.2015.326.

Li, W., Zhang, Z. and Liu, Z. (2010) 'Action Recognition Based on A Bag of 3D Points.pdf', *Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on*, 2010, pp. 9–14. doi: 10.1109/CVPRW.2010.5543273.

Liu, S., Chen, C. and Kehtarnavaz, N. (2016) 'A computationally efficient denoising and hole-filling method for depth image enhancement', 9897, p. 98970V. doi: 10.1117/12.2230495.

Moeslund, T. B., Hilton, A. and Krüger, V. (2006) 'A survey of advances in vision-based human motion capture and analysis', *Computer Vision and Image Understanding*, 104(2–3 SPEC. ISS.), pp. 90–126. doi: 10.1016/j.cviu.2006.08.002.

Ning, H. N. H. *et al.* (2008) 'Discriminative learning of visual words for 3D human pose estimation', *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8. doi: 10.1109/CVPR.2008.4587534.

Ortiz-catalan, M. *et al.* (2014) 'Emerging therapies in Neurorehabilitation', *Emerging Therapies in Neurorehabilitation Biosystems & Biorobotics Volume 4, 2014, pp 249-265*, 4, pp. 249–265. doi: 10.1007/978-3-642-38556-8.

Oyama, M. *et al.* (2017) 'Two-stage Model Fitting Approach for Human Body Shape Estimation from a Single Depth Image', (Figure 1), pp. 8–11.

Park, S. *et al.* (2017) 'Accurate and Efficient 3D Human Pose Estimation Algorithm Using Single Depth Images for Pose Analysis in Golf', *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 105–113. doi: 10.1109/CVPRW.2017.19.

Pekelny, Y. and Gotsman, C. (2008) 'Articulated object reconstruction and markerless motion capture from depth video', *Computer Graphics Forum*, 27(2), pp. 399–408. doi: 10.1111/j.1467-8659.2008.01137.x.

Plagemann, C. *et al.* (2010) 'Real-time Identification and Localization of Body parts from depth images', *Proceedings - IEEE International Conference on Robotics and Automation*, pp. 3108–3113. doi: 10.1109/ROBOT.2010.5509559.

Poppe, R. (2007) ‘Vision-based human motion analysis: An overview’, *Computer Vision and Image Understanding*, 108(1–2), pp. 4–18. doi: 10.1016/j.cviu.2006.10.016.

Rius, I. *et al.* (2009) ‘Action-specific motion prior for efficient Bayesian 3D human body tracking’, *Pattern Recognition*. Elsevier, 42(11), pp. 2907–2921. doi: 10.1016/j.patcog.2009.02.012.

Rosales, R. and Sclaroff, S. (2006) ‘Combining generative and discriminative models in a framework for articulated pose estimation’, *International Journal of Computer Vision*, 67(3), pp. 251–276. doi: 10.1007/s11263-006-5165-4.

Sarafianos, N. *et al.* (2016) ‘3D Human pose estimation: A review of the literature and analysis of covariates’, *Computer Vision and Image Understanding*. Elsevier Inc., 152, pp. 1–20. doi: 10.1016/j.cviu.2016.09.002.

Satpathy, M., Siebel, N. T. and Rodriguez, D. (2004) ‘Assertions in Object Oriented Software Maintenance: Analysis and Case Study’, *20th Ieee International Conference on Software Maintenance, Proceedings*, (September), pp. 124–133. doi: 10.1109/icsm.2004.1357797.

Schwarz, L. A. *et al.* (2012) ‘Human skeleton tracking from depth data using geodesic distances and optical flow’, *Image and Vision Computing*, pp. 217–226. doi: 10.1016/j.imavis.2011.12.001.

Sedai, S., Bennamoun, M. and Huynh, D. (2010) ‘Localized fusion of Shape and Appearance features for 3D Human Pose Estimation’, *Procedings of the British Machine Vision Conference 2010*, p. 51.1-51.10. doi: 10.5244/C.24.51.

Sedai, S., Bennamoun, M. and Huynh, D. Q. (2013) ‘A Gaussian Process Guided Particle Filter for Tracking 3D Human Pose in Video’, *Image Processing, IEEE Transactions on*, 22(11), pp. 4286–4300. doi: 10.1109/TIP.2013.2271850.

Shotton, J., Girshick, R., *et al.* (2013) ‘Efficient human pose estimation from single depth images’, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(12), pp. 2821–2840. doi: 10.1109/TPAMI.2012.241.

Shotton, J., Fitzgibbon, A., *et al.* (2013) ‘Real-time human pose recognition in parts from single depth images’, *Studies in Computational Intelligence*, 411, pp. 119–135. doi:

10.1007/978-3-642-28661-2-5.

Sigal, L. *et al.* (2012) ‘Loose-limbed people: Estimating 3D human pose and motion using non-parametric belief propagation’, *International Journal of Computer Vision*, 98(1), pp. 15–48. doi: 10.1007/s11263-011-0493-4.

Sigal, L., Balan, A. O. and 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(1–2), pp. 4–27. doi: 10.1007/s11263-009-0273-6.

Sminchisescu, C. (2011) ‘Estimation algorithms for ambiguous visual models : Three Dimensional Human Modeling and Motion Reconstruction in Monocular Video Sequences par’.

Suma, E. A. *et al.* (2011) ‘FAAST: The flexible action and articulated skeleton toolkit’, *Proceedings - IEEE Virtual Reality*, pp. 247–248. doi: 10.1109/VR.2011.5759491.

Tekin, B. *et al.* (2015) ‘Direct Prediction of 3D Body Poses from Motion Compensated Sequences’. doi: 10.1109/CVPR.2016.113.

Tian, Y. *et al.* (2013) ‘Canonical locality preserving Latent Variable Model for discriminative pose inference’, *Image and Vision Computing*, 31(3), pp. 223–230. doi: 10.1016/j.imavis.2012.06.009.

Veronese, A. *et al.* (2017) ‘Probabilistic Mapping of Human Visual Attention from Head Pose Estimation’, *Frontiers in Robotics and AI*, 4(October), pp. 1–11. doi: 10.3389/frobt.2017.00053.

Wang, C. *et al.* (2014) ‘Robust Estimation of 3D Human Poses from a Single Image’, (013). doi: 10.1109/CVPR.2014.303.

Wei, X., Zhang, P. and Chai, J. (2012) ‘Accurate realtime full-body motion capture using a single depth camera’, *ACM Transactions on Graphics*, 31(6), p. 1. doi: 10.1145/2366145.2366207.

Weiss, A., Hirshberg, D. and Black, M. J. (2011) ‘Home 3D Body Scans from Noisy Image and Range Data † Perceiving Systems Dept ., Max Planck Institute for Intelligent Systems , T’, *Work*, pp. 1951–1958. doi: 10.1109/ICCV.2011.6126465.

Wigdor, D. and Wixon, D. (2011) *Brave NUI World: Designing Natural User Interfaces for Touch and Gesture*, *Brave NUI World Designing Natural User Interfaces for Touch and Gesture*. doi: 10.1016/B978-0-12-382231-4.X0001-9.

Wouterse, P. (2015) 'Using Facial Feature Recognition and Head Tracking to Control Games', pp. 1–41.

Wu, D., Zhu, F. and Shao, L. (2012) 'One shot learning gesture recognition from RGBD images', *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pp. 7–12. doi: 10.1109/CVPRW.2012.6239179.

Yasin, H. *et al.* (2016) 'A Dual-Source Approach for 3D Pose Estimation from a Single Image', *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4948–4956. doi: 10.1109/CVPR.2016.535.

Ye, G. *et al.* (2012) 'Performance capture of interacting characters with handheld kinects', *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 7573 LNCS(PART 2), pp. 828–841. doi: 10.1007/978-3-642-33709-3_59.

Ye, M. *et al.* (2013) 'A survey on human motion analysis from depth data', *Lecture Notes in Computer Science (Time-of-Flight and Depth Imaging. Sensors, Algorithms, and Applications)*, 8200 LNCS, pp. 149–187. doi: 10.1007/978-3-642-44964-2_8.

Zheng, Y. and Yamane, K. (2015) 'Human motion tracking control with strict contact force constraints for floating-base humanoid robots', *IEEE-RAS International Conference on Humanoid Robots*, 2015–Febru(February), pp. 34–41. doi: 10.1109/HUMANOIDS.2013.7029952.

Zuffi, S. and Black, M. J. (2015) 'The Stitched Puppet : A Graphical Model of 3D Human Shape and Pose', pp. 3537–3546.

Anexo

Año	*Trabajo*	*Datos Entrada*	*Vistas*	*Tipo de Método*	*Modelo*	*Algoritmos principales*	*Conjuntos de datos*
2017	A Multi-view RGB-D Approach for Human Pose Estimation in Operating Rooms (Kadkhodamohammadi *et al.*, 2017)	Imágenes RGB y de profundidad	Multivista	Discriminativo	PSM	Red convolucional	Conjunto propio
	Accurate and Efficient 3D Human Pose Estimation Algorithm using Single Depth Images for Pose Analysis in Golf (Park *et al.*, 2017)	Imagen de profundidad	Monocular	Discriminativo	PSM	Arboles de regresión	Conjunto propio
	PoseTrack: Joint Multi-Person Pose Estimation and Traquing (Iqbal, Milan y Gall, 2017)	Secuencia de imágenes RGB	Monocular	Generativo	PSM	*Tracking*	Conjunto propio
	Two-stage Model Fitting Approach for Human Body Shape Estimation from a Single Depth Image (Oyama *et al.*, 2017)	Imagen de profundidad	Monocular	Generativo	*Stitched Puppet*	Editor de superficies de Laplace	Conjunto propio
2016	A Dual-Source Approach for 3D Pose Estimation from a Single Image (Yasin *et al.*, 2016)	Imagen RGB + captura de movimiento 3D	Monocular	Discriminativo	PSM	Arboles de regresión	CMU, Human3.6M, HumanEva-I
2015	Depth-Images-Based Pose Estimation Using Regression Forests and Graphical Models (He *et al.*, 2015)	Imagen de profundidad	Monocular	Híbrido	PSM	Arboles de regresión con PSM	Conjunto propio, Standford Dataset of real Depth Images
	Direct prediction of 3D Body Poses from Motion Compensated Sequences (Tekin *et al.*, 2015)	Secuencia de imágenes RGB	Monocular	Discriminativo	PSM	Redes convolucionales + KRR y KDE	Human3.6M, HumanEva, KTH

	Efficient ConvNet-based Marker-less Motion Capture in Generaal Scenes with a Low Number of Cameras (Elhayek *et al.*, 2015)	Secuencia de imágenes RGB	Multivista	Híbrido	PSM	Red convolucional + *Tracking*	Conjunto propio
	Maximum-Margin Structured Learning with Deep Networks for 3D Human Pose Estimation (Li, Zhang y Chan, 2015)	Imagen RGB	Monocular	Discriminativo	PSM	Red convolucional	Human3.6M
	Multiple Human Pose Estimation with Temporally Consistent 3D Pictorial Structures (Belagiannis *et al.* 2015)	Secuencia de imágenes RGB	Multivista	Generativo	3DPSM	*Tracking*	Campus, Shelf
2014	3D Human Pose Estimation from Monocular Images with Deep Convolutional Neural Network (Li y Chan, 2014)	Imagen RGB	Monocular	Discriminativo	PSM	Red convolucional	Human3.6M
	Depth Sweep Regession Forest for Estimating 3D Human Pose from Images (Kostricov y Gall, 2014)	Imagen RGB	Monocular	Híbrido	PSM	Arboles de regresión	Human3.6M, HumanEva-I
	Robust Estimation of 3D Human Poses from a Single Image (Wang *et al.*, 2014)	Imagen RGB	Monocular	Discriminativo	PSM	Pose 2D + parámetros de la cámara	CMU, HumanEva, UvA 3D
2013	A Gaussian Process Guided Particle Filter for Tracking 3D Human Pose in Video (Sedai, Bennamoun y Huynh, 2013)	Secuencia de imágenes RGB	Monocular	Híbrido	PSM	Filtrado de partículas + *Tracking*	HumanEva
	Efficient Human Pose Estimation from Single Depth Images (Shoton, Girshick *et al.*, 2013)	Imagen de profundidad	Monocular	Discriminativo	---	Árboles de decisiones aleatorias	MSRC-5000
	Multi-view Pictorial Structures for 3D Human Pose Estimation (Amin *et al.*, 2013)	Imágenes RGB	Multivista	Generativo	PSM	Combinación de PSMs + Triangulación	HumanEva-I, MPII Cooking
	Real-Time Human Pose Recognition in Parts from Single Depth Images (Shoton *et al.*, 2013)	Imagen de profundidad	Monocular	Discriminativo	---	Árboles de decisiones aleatorias + detección de partes del cuerpo	Conjunto propio

2012	Accurate Realtime Full-body Motion Capture Using a Single Depth Camera (Wei, Zhang y Chai, 2012)	Secuencia de imágenes de profundidad	Monocular	Híbrido	PSM	Arboles de decisiones aleatorias + *Tracking*	Conjunto propio
2011	Efficient Regression of General-Activity Human Poses from Depth Images (Girshick *et al.*, 2011)	Imagen de profundidad	Monocular	Discriminativo	PSM	Arboles de regresión	MSRC-5000, Standford Dataset of real Depth Images
2010	Fast Local and Global Similarity Searches in Large Motion Capture Databases (Krüger *et al.*, 2010)	Secuencia de imágenes de profundidad	Monocular	Híbrido	PSM	Extremos geodésicos + *Tracking*	CMU, HDM05
	Monocular 3D Pose Estimation and Tracking by Detection (Andriluka *et al.*, 2010)	Secuencia de imágenes RGB	Monocular	Generativo	PSM	*Tracking*	HumanEva-II
	Real-time Identification and Localization of Body Parts from Depth Images (Plagemann *et al.*, 2010)	Imagen de profundidad	Monocular	Discriminativo	---	Identificación de extremos geodésicos	Conjunto propio
	Real Time Motion Capture Using a Single Time-Of-Flight Camera (Ganapathi *et al.*, 2010)	Secuencia de imágenes de profundidad	Monocular	Híbrido	PSM + Malla de superficie	Extremos geodésicos + *tracking*	Conjunto propio
2009	Action-specific Motion prior for efficient Bayesian 3D human body tracking (Rius *et al.*, 2009)	Secuencia de imágenes RGB	Monocular	Generativo	PSM	*Tracking*	HumanEva-I
	Fusion of 2D and 3D sensor data for articulated body tracking	Secuencia de imágenes RBG y de profundidad	Monocular	Generativo	3DPSM	*Tracking*	Conjunto propio