Escuela
Politécnica
Superior

Reconstrucción del cuerpo humano mediante imágenes RGB-D y técnicas de visión artificial y deep learning

Grado en Ingeniería Informática

Trabajo Fin de Grado

Autora:

Camelia Beltrá García

Tutores:

Jorge Azorín López

Andrés Fuster Guilló

Universitat d'Alacant
Universidad de Alicante

Reconstrucción del cuerpo humano mediante imágenes RGB-D y técnicas de visión artificial y deep learning

Subtítulo del proyecto

Autora

Camelia Beltrá García

Tutores

Jorge Azorín López

Andrés Fuster Guilló

TECNOLOGIA INFORMATICA Y COMPUTACION

Grado en Ingeniería Informática

Escuela Politécnica Superior

Universitat d'Alacant
Universidad de Alicante

ALICANTE, Junio 2022

Resumen

La visión artificial se encuentra en constante evolución en parte motivada por los avances de la Inteligencia Artificial y la optimización basada en Deep Learning.

El Deep Learning es un enfoque de aprendizaje automático no supervisado (es decir, son necesarios datos de entrenamientos, pero estos no requieren ser etiquetados) que se asemeja o se fundamenta en el funcionamiento del sistema neurológico humano.

La visión artificial es una disciplina científica que incluye métodos para adquirir, procesar y analizar imágenes del mundo real con el fin de producir información que pueda ser tratada por una máquina, la principal finalidad de esta tecnología es dotar a la máquina de ”ojos” para ver que ocurre en el mundo real, y así poder tomar decisiones para automatizar cualquier proceso.

La razón principal que me ha llevado a la realización de este trabajo ha sido el uso de ambas tecnologías comentadas anteriormente, dado están cada vez más en auge además de la propia curiosidad sobre el aprendizaje y funcionamiento de estas.

Este trabajo tiene como finalidad el estudio de la visión artificial dentro de un marco dietético nutricional de la obesidad aprovechando las tecnologías para abordar este problema proponiendo una solución de la representación del cuerpo humano 3D con el uso único de imágenes y realizar unas comparativas sobre los modelos conseguidos para comprobar la veracidad de los datos obtenidos a partir de una red neuronal.

Agradecimientos

Este trabajo no habría sido posible sin el apoyo de mis tutores como de Nahuel, su paciencia y disposición hacia mis dudas han sido clave durante este proyecto.

También quiero agradecer a mis amigos, que por muy lejos que estén han estado cerca de mí apoyándome y animándome para terminar mis estudios en los momentos más complejos.

Agradecer a mi familia, que siempre han tratado de darme soluciones a mis problemas.

No puedo terminar estos agradecimientos sin mencionar a mis compañeros de la Universidad de Alicante; Luis, Sandra, Edgar, Dani y el resto de personas que han estado a mi lado, gracias por los años que he estado a vuestro lado y por toda la ayuda y apoyo que nos hemos brindado, estoy muy agradecida de haberos tenido a mi lado durante mis años de carrera.

Es a ellos a quien dedico este trabajo.

A mi abuela Carmen, que no ha podido verme terminar esta etapa.

Índice general

1 Introducción **1**

1.1 Motivación y contexto 1

1.2 Estado del arte 4

1.3 Objetivos 8

Bibliografía **9**

1 Introducción

1.1 Motivación y contexto

Mi interés por la visión artificial se desarrolló durante el tercer año de carrera, donde me presenté junto a la Universidad de Alicante a un proyecto llamado Vodafone Campus Lab[Vodafone (2020)], donde te proponen unos problemas a resolver y nosotros decidimos plantear una solución con tecnologías como la visión artificial, además esta propuesta requería de procesos como la adquisición del cuerpo 3D y reconocimiento de este, entre otros. Aquí fue cuando busqué propuestas de trabajo de fin de grado (TFG) similares o sobre este tema, porque me quedé con la curiosidad de llevarlo a cabo y porque quería saber, aprender e investigar más sobre la adquisición del objetos 3D a partir de imágenes, cuando vi las propuestas fue cuando conocí el proyecto de investigación [Tech4Diet (2019)].

El proyecto de investigación Tech4Diet cuenta con el apoyo de la Agencia Estatal de Investigación (AEI) y del Fondo Europeo de Desarrollo Regional (FEDER) con referencia "TIN2017-89069-R" perteneciente al programa Retos 2017 en el que su investigador jefe es Jorge Azorín. En este proyecto se busca facilitar el estudio de la evolución morfológica ocasionada por tratamientos de obesidad. Hoy en día, estos tratamientos son muy costosos pero a su vez muy necesarios, ya que los problemas de obesidad o sobrepeso pueden ocasionar enfermedades crónicas como la hipertensión, diabetes tipo II, cáncer. También pueden ocasionar enfermedades patológicas neurodegenerativas como el Alzheimer o demencias [Fuster-Guilló y cols. (2020)]

El sistema utilizado para esta finalidad, dispone de una red de cámaras RGB-D que obtienen un modelo 3D del cuerpo del paciente. El proceso de obtención de un modelo 3D se realiza en diferentes sesiones médicas, lo que permite una visualización real de la evolución del cuerpo del paciente. Para que el paciente pueda visualizar su progreso, no solo dispondrá de una aplicación de ordenador, sino que también podrá visualizar los diferentes modelos de su cuerpo mediante unas gafas de realidad virtual. La realidad virtual tiene como finalidad incrementar la adherencia del usuario al tratamiento. Además, podemos encontrar desarrollos tecnológicos como "Google Cardboard" que nos permiten convertir cualquier teléfono móvil en unas gafas de realidad virtual sin necesidad de realizar un gasto de dinero elevado. A parte de la visualización, sobre este modelo 3D obtenido con las cámaras se pueden realizar mediciones de diferentes partes del cuerpo a niveles de 1D, 2D y 3D.

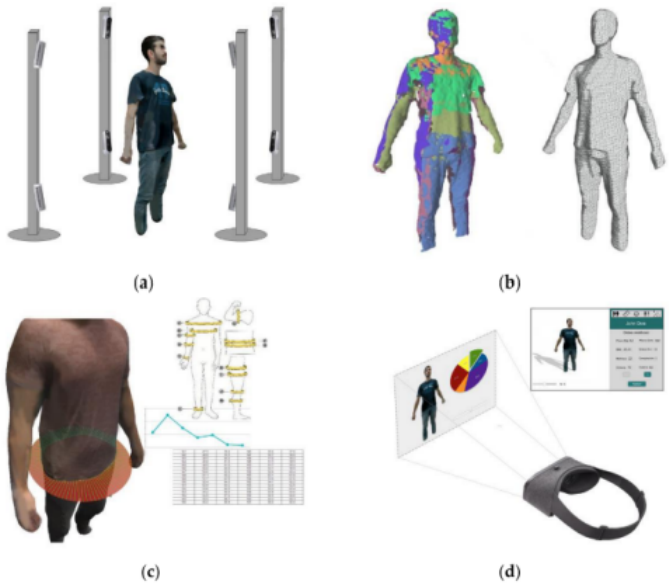

Figura 1.1: Captación del modelo del cuerpo mediante el sistema de cámaras (a). Nubes de puntos sin textura generada (b). Mediciones del cuerpo (c). Visualización de los resultados en la aplicación y con gafas de realidad virtual [García-D'Urso y cols. (2021)]

Este trabajo cuenta como objetivo principal la reconstrucción del cuerpo a partir

de imágenes dado que esto permite que los proyectos mencionados anteriormente se puedan exportar de una manera más económica, por lo tanto para que todo esto sea efectivo las mediciones han de corresponderse a las reales, por lo que el siguiente objetivo será comprobar, calcular y comparar los resultados obtenidos mediante el uso de diferentes cámaras RGB-D con nuestros resultados generados a partir de un modelo 3D generado con una sola imagen del cuerpo de la persona que será procesado en una red que alinea los píxeles de las imágenes con su correspondiente contexto global del modelo 3D [Saito y cols. (2019)].

1.2 Estado del arte

La reconstrucción del cuerpo humano 3D es un tema amplio abordado de diferentes maneras, por una parte tenemos la visión por computador que nos permite entender la información visual del entorno capturada a través de las cámaras [Z. Zhang (2016)], por otro lado tenemos la visión artificial, por este lado están apareciendo avances recientes en la estimación del cuerpo humano 3D basado en imágenes que han sido impulsados por la mejora significativa en el poder de representación que ofrecen las redes neuronales profundas. [Saito y cols. (2020)]

Se ha llevado a cabo una revisión del estado del arte que abordan las diferentes formas de obtener resultados respecto a la reconstrucción del cuerpo humano.

Por un lado, recientes trabajos abordados mediante el uso de la visión por computador utilizan cámaras RGB-D calibradas [García-D'Urso y cols. (2021)]. El trabajo abordado por García D'Urso propone el uso de dispositivos RGB-D (como Microsoft Kinect o Intel RealSense), debido a que integran sensores de color y profundidad, y utilizan tecnologías de profundidad como luz estructurada, ToF (Time of Flight, sensor que mide distancias utilizando el tiempo que usan los fotones en viajar entre dos puntos) o activa estereoscópica. En este proyecto aparte de la obtención del modelo, también han tratado otros puntos como la visualización 3D del cuerpo utilizando la realidad virtual, además de que pueden obtener las medidas de volúmenes seleccionados del cuerpo humano. Con esta propuesta lo que llevan a cabo es una red de cámaras RGB-D de 13 cámaras ubicadas en una cabina donde la persona se ha de colocar durante 4 segundos como se puede ver en la figura 1.2, esta red de cámaras se ha de calibrar de manera que cada una de ellas quede calibrada intrínsicamente y extrínsecamente para cada cámara y multi-cámara para así estimar la posición relativa de cada sensor en la red, una vez la red de cámaras se encuentren calibradas sigue el proceso de la obtención del modelo 3D con textura que consta de 5 fases principales, que se ven en la figura 1.1: adquisición, pre-procesado, unión de vistas, generación de la malla y

proyección de la textura.

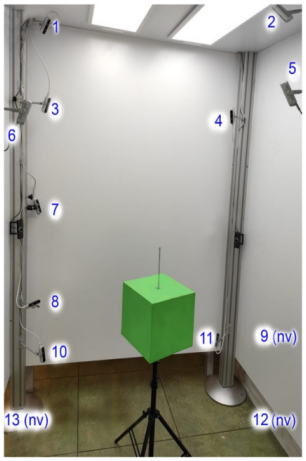

Figura 1.2: Set-up de red de 13 cámaras RGB-D. "nv" indica que la cámara no está visible en la imagen [García-D'Urso y cols. (2021)].

Por otro lado nos encontramos recientes trabajos abordados mediante el uso de visión artificial y deep learning, en estos a diferencia del anterior no se necesita una cabina formada por una red de cámaras. Todos estos trabajos tienen el mismo objetivo de producir resultados naturales y bien alineados [H. Zhang y cols. (2021)], en concreto han habido dos paradigmas investigados principalmente, por un lado tenemos los métodos basados en la optimización que ajustan explícitamente los modelos a las imágenes 2D [Bogo y cols. (2016)], y por otro lado los basados en modelos de regresión que sugieren directamente predecir los parámetros del modelo a partir de las imágenes [H. Zhang y cols. (2021)].

Bogo desarrolla un método para automáticamente estimar la pose y forma 3D del cuerpo humano a partir de una imagen, para ello siguen dos pasos, primero estiman los 2D joints (uniones, en este caso sería por ejemplo la rodilla que une la parte inferior y superior de la pierna), esto lo consiguen usando una red neuronal convolucional (CNN) llamado DeepCut [Pishchulin y cols. (2015)], esta red es buena estimando la pose 2D pero no bueno con la pose 3D por lo que el siguiente paso es el que estima la pose y la forma 3D partiendo de los 2D joints utilizando el modelo SMPL [Loper y cols. (2015)]

encanjando dentro de un paradigma clásico bottom up de estimación (CNN) seguida de una verificación descendente (modelo generativo) como se puede ver en la figura 1.3

Figura 1.3: Ejemplos. forma y pose 3D estimada mediante el método desarrollado por Bogo[Bogo y cols. (2016)] usando las fotos de Leeds Sports Pose Dataset[Johnson y Everingham (2010)]. Se ve la imagen original a la izquierda, luego el modelo comparado en la foto en el centro y por último a la derecha el modelo 3D desde otro punto de vista.

Los métodos basados en la optimización como el de Bogo ajustan explícitamente los modelos a las imágenes, lo que produce resultados con buena precisión de los alineamientos entre malla-imagen, pero tienden a ser lentos y sensibles al inicio[H. Zhang y cols. (2021)].

Por otro lado nos encontramos los modelos basados en regresión como la propuesta de H. Zhang, estos modelos se basan en los parámetros existentes en las imágenes, estos modelos están demostrando resultados prometedores pero aún nos encontramos errores principalmente de alineamientos entre malla e imagen. La propuesta de H. Zhang es principalmente para tratar de resolver este problema, propone y diseña PyMAF (Pyramidal Mesh Alignment Feedback). La idea central de este enfoque es corregir las desviaciones paramétricas de manera explícita y progresiva basado en el estado del alineamiento, en PyMAF el alineamiento se realiza extrayendo las características espaciales respecto a la proyección 2D de la malla estimada y luego retroalimenta los regresores para que actualicen los parámetros.

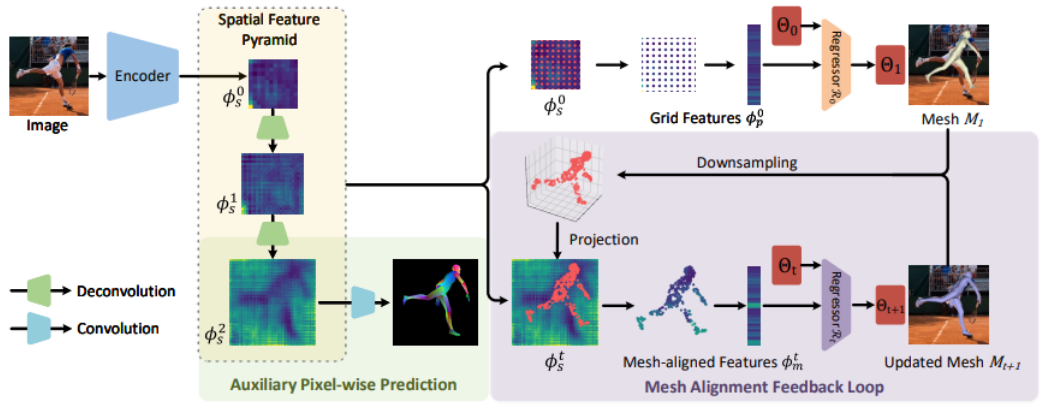

Figura 1.4: Descripción general de la propuesta PyMAF. PyMAF aprovecha una Feature Pyramid y permite un circuito de retroalimentación de alineación en la red. Dada una predicción del modelo, los alineamientos de la malla son extraídos de características de una resolución más fina y retroalimentan el regresor para así actualizar la malla. [H. Zhang y cols. (2021)]

Con esta investigación del estado del arte podemos observar que hay diferentes maneras de realizar la reconstrucción del cuerpo humano, dentro de cada manera hay varios trabajos parecidos dado que unos se retroalimentan de otros.

Los trabajos realizados mediante visión artificial y redes neuronales están más enfocados en general a las poses de las personas, dado que comenzó inicialmente en ello, pero lo que nosotros queremos y buscamos es que sean capaces de reconstruir de forma fiel el cuerpo dado que se busca usar esto en un ámbito dietético-nutricional de la obesidad, por lo que es muy importante que las medidas sean correctas.

1.3 Objetivos

Este trabajo esta enmarcado dentro de un proyecto de investigación, y por lo tanto comparte el mismo objetivo general, proporcionar un sistema capaz de estudiar la evolución del cuerpo humano con el paso del tiempo mediante técnicas de visión 3D.

Para llegar al objetivo general es necesario disponer de una buena obtención del modelo 3D. Para ello, tenemos los siguientes objetivos:

- **Objetivo 1: Obtención del modelo 3D a partir de una imagen usando PIFu como red.**

  Para ello se realizarán tareas como:

  - Estudio de los diferentes parámetros para el entrenamiento y prueba de la red.
  - Obtención de imagen del cuerpo y generación de máscara de la imagen.
  - Estudio de los resultados obtenidos.

- **Objetivo 2: Comparativa del modelo 3D obtenido con el modelo obtenido por el proyecto Tech4Diet.**

  Para este objetivo se tendrán que realizar las siguientes tareas específicas:

  - Estudio sobre las diferentes formas de realizar el cálculo de las diferencias obtenidas.
  - Cálculo de distancias usando Hausdorff Distance
  - Cálculo de distancias usando Chamfer distance

Bibliografía

Bogo, F., Kanazawa, A., Lassner, C., Gehler, P. V., Romero, J., y Black, M. J. (2016). Keep it SMPL: automatic estimation of 3d human pose and shape from a single image. *CoRR*, *abs/1607.08128*. Descargado de `http://arxiv.org/abs/1607.08128`

Fuster-Guilló, A., Jorge Azorín-López, M. S.-C., Castillo-Zaragoza, J. M., Garcia-D'Urso, N., y Fisher, R. B. (2020). RGB-D based framework to Acquire, Visualize and Measure the Human Body for Dietetic Treatments..

García-D'Urso, N., Galán-Cuenca, A., Manchón-Pernis, C., Furster-Guilló, A., y Azorín-López, J. (2021). Arquitectura de visión 3D para medición y visualización del cuerpo humano..

Johnson, S., y Everingham, M. (2010). Clustered pose and nonlinear appearance models for human pose estimation. En *Proceedings of the british machine vision conference* (pp. 12.1–12.11). BMVA Press. (doi:10.5244/C.24.12)

Kanazawa, A., Black, M. J., Jacobs, D. W., y Malik, J. (2017). End-to-end recovery of human shape and pose. *CoRR*, *abs/1712.06584*. Descargado de `http://arxiv.org/abs/1712.06584`

Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., y Black, M. J. (2015). Smpl: a skinned multi-person linear model. *ACM Trans. Graph.*, *34*, 248:1-248:16.

Pishchulin, L., Insafutdinov, E., Tang, S., Andres, B., Andriluka, M., Gehler, P. V., y Schiele, B. (2015). Deepcut: Joint subset partition and labeling for multi person

pose estimation. *CoRR, abs/1511.06645*. Descargado de http://arxiv.org/abs/1511.06645

Saito, S., Huang, Z., Natsume, R., Morishima, S., Kanazawa, A., y Li, H. (2019). Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. *CoRR, abs/1905.05172*. Descargado de http://arxiv.org/abs/1905.05172

Saito, S., Simon, T., Saragih, J., y Joo, H. (2020). PIFuHD: Multi-Level Pixel-Aligned Implicit Function for High-Resolution 3D Human Digitization..

Tech4Diet. (2019). Project tin2017-89069-r spanish state research agency (aei). 4d modelling and visualization of the human body to improve adherence to dietetic-nutritional intervention of obesity.. (http://tech4d.dtic.ua.es/tech4d// (accessed Jun. 2022))

Vodafone. (2020). Vodafone campus lab, el programa que te enseña a desarrollar soluciones a grandes problemas sociales en 4 meses, con la ayuda de los mejores expertos de vodafone.. (https://vodafonecampuslab.es (accessed Jun. 2022))

Zhang, H., Tian, Y., Zhou, X., Ouyang, W., Liu, Y., Wang, L., y Sun, Z. (2021). 3d human pose and shape regression with pyramidal mesh alignment feedback loop. *CoRR, abs/2103.16507*. Descargado de https://arxiv.org/abs/2103.16507

Zhang, Z. (2016). Camera calibration: a personal retrospective. Machine Vision and Applications.. (https://doi.org/10.1007/s00138-016-0809-z)