

CARRERA DE ESPECIALIZACIÓN EN INTELIGENCIA ARTIFICIAL

MEMORIA DEL TRABAJO FINAL

Yoga Pose Estimation

Autor:
Ing. Juan Ignacio Ribet

Director:
Ing. Juan Pablo Pizarro (Globant)

Jurados:
Nombre del jurado 1 (pertenencia)
Nombre del jurado 2 (pertenencia)
Nombre del jurado 3 (pertenencia)

*Este trabajo fue realizado en la ciudad de Suipacha,
entre mayo de 2023 y diciembre de 2023.*

Resumen

La presente memoria describe el desarrollo de un detector de postura de yoga o asanas para la empresa Globant. Este trabajo pretende obtener el mínimo producto viable de una aplicación para entrenamiento de yoga, en donde se detecte la postura mediante algoritmos de aprendizaje de máquina y visión por computadora, y se le de soporte al usuario para realizarlas de forma correcta.

Para su desarrollo fueron fundamentales los conocimientos adquiridos en la carrera, en especial los vistos en las materias de visión por computadora con el uso indispensable de la librería OpenCV como también los conceptos de aprendizaje de máquina, análisis de datos y la comprensión del mecanismo de funcionamiento de las de las redes neuronales convolucionales.

Índice general

| | |
|---|-----------|
| Resumen | I |
| 1. Introducción general | 1 |
| 1.1. Motivación | 1 |
| 1.2. Objetivos y alcance | 2 |
| 1.3. Introducción a la estimación de pose | 2 |
| 1.3.1. Evolución de la estimación de pose en visión por computadora | 3 |
| 1.3.2. Aplicaciones | 4 |
| 1.4. Estado del arte | 4 |
| 2. Introducción específica | 7 |
| 2.1. Técnicas de visión por computadora | 7 |
| 2.2. Redes neuronales con capas convolucionales. | 8 |
| 2.3. Detección de objetos | 10 |
| 2.4. Modelado y seguimiento del cuerpo humano. | 11 |
| 2.5. Descripción del modelo MediaPipe (<i>BlazePoze</i>) | 13 |
| 2.5.1. Arquitectura del modelo y diseño del pipeline | 13 |
| Bibliografía | 17 |

Capítulo 1

Introducción general

1.1. Motivación

La motivación para realizar este trabajo tiene fundamento en varias aristas, como adoptar las nuevas tecnologías en visión por computadora y aprendizaje de máquina, para la práctica de una disciplina tan ampliamente realizada a nivel global y con beneficios probados para el bienestar y salud de las personas. Gracias al avance en los modelos de algoritmos de aprendizaje profundo y el acceso a grandes conjuntos de datos, la capacidad de las computadoras para comprender y procesar imágenes ha mejorado significativamente. Esto ha llevado a mejoras sustanciales en reconocimiento facial, detección de objetos, segmentación de imágenes y en particular para el interés de este trabajo, estimación de poses humanas.

Por otro lado, se ha identificado como un motor de este trabajo la transformación en las prácticas de yoga que fueron fruto del impacto de la pandemia de Covid-19 [1]. Con la implementación del aislamiento y el distanciamiento social, se observó un aumento en el trabajo remoto como modalidad laboral predominante. Como resultado, muchas personas experimentaron una vida más sedentaria combinada con la imposibilidad de participar en actividades físicas fuera del hogar [2]. En respuesta a la situación vivenciada en este período, se vio un aumento de interés en realizar ejercicios de forma virtual en diferentes plataformas. Como resultado de estas transformaciones, se observó un aumento en la búsqueda de términos relacionados con “yoga” en motores de búsqueda como Google y plataformas de video como YouTube [2] como se ilustra en el gráfico de la figura 1.1 y refleja el creciente interés en realizar esta disciplina.

Los acontecimientos nombrados anteriormente permiten desarrollar, con base a los conocimientos adquiridos en la formación de la Especialización en Inteligencia Artificial, una interfaz que permita corregir de manera autónoma las prácticas de yoga. En particular aquellas que se realizan en solitario y así, evitar sufrir malestares que sean fruto de una mala aplicación de las posturas. Como así también poder asegurar un avance y mejora en las mismas.

Globant S.A., encuentra la transformación descripta anteriormente como una oportunidad de negocio para desarrollar un software, basado en tecnologías innovadoras como visión por computadora, que permita mejorar la experiencia de usuario en la práctica autónoma de esta disciplina.

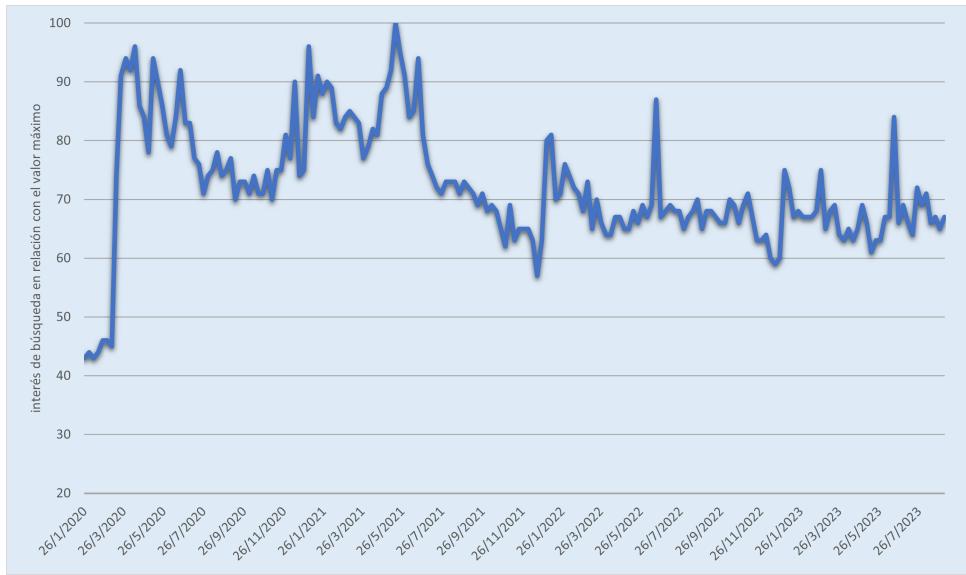


FIGURA 1.1. Busquedas de “yoga” en YouTube desde inicio de 2020 a la fecha.

1.2. Objetivos y alcance

El proyecto tiene un objetivo primario: el desarrollo de un prototipo de aplicación, que cuente con la función de detectar las posturas de yoga, utilizando las herramientas disponibles en visión por computadora, aprendizaje de maquina y análisis de datos.

El sistema deberá poder ser utilizado en dispositivos de bajos recursos computacionales, que a su vez dispongan de cámara web para la obtención de los videos. Estas dos condiciones son necesarias pues la detección tiene que ser realizada en tiempo real. Complementando la detección el modelo buscará mostrar referencias visuales al usuario que lo ayuden a lograr la postura correctamente como se puede ver en el ejemplo de la figura 1.2. Una vez finalizada la práctica, el usuario tendrá acceso a un informe detallado del ejercicio realizado. La devolución se presentarán de forma clara y concluyente y así poder mejorar la realización de las asanas.

Este prototipo implica el desarrollo de una interfaz simple y funcional. Para ser utilizada por usuarios sin conocimientos técnicos. La aplicación, a su vez, tomará como premisa que la persona cuenta con conocimientos básicos de yoga.

Para finalizar, es importante aclarar que las posturas a detectar estarán limitadas a una cantidad que permita mostrar el mecanismo y funcionalidad de la aplicación. Siendo al mismo tiempo una prioridad poder cumplir con el presupuesto de tiempo y recursos, teniendo en cuenta la raíz académica del proyecto.

1.3. Introducción a la estimación de pose

La estimación de pose es una técnica de visión por computadora, que identifica las articulaciones claves del cuerpo de un ser humano en imágenes y videos, para comprender su postura. Si bien la estimación de la pose también se puede aplicar

a diversos objetos, existe un interés particular en la pose humana debido a su amplia gama de aplicaciones prácticas y su impacto social.

La resolución del problema de estimación de la postura humana articulada a partir de imágenes tiene una extraordinaria dificultad. Los algoritmos tienen que lidiar con una gran cantidad de poses, grandes cambios en la apariencia, oclusiones parciales y la presencia de varias personas [3].

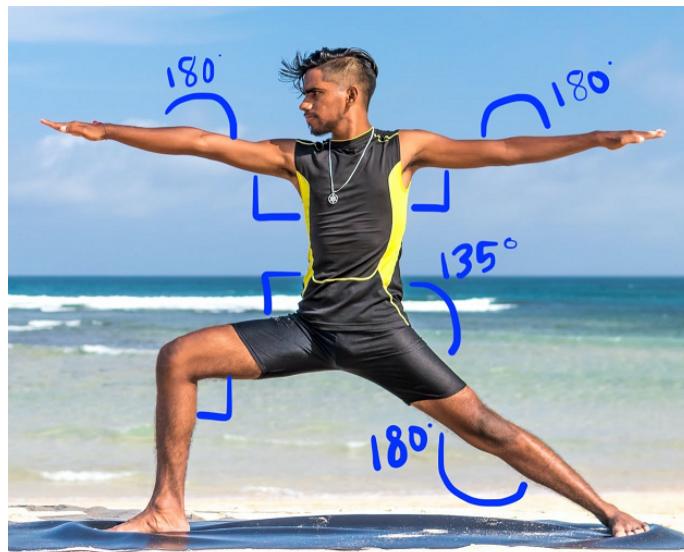


FIGURA 1.2. Ejemplo de indicaciones visuales en postura de yoga.

1.3.1. Evolución de la estimación de pose en visión por computadora

La investigación sobre la estimación de la pose comenzó con el surgimiento de la visión por computadora como campo a finales de los años 1960 y principios de los 1970. Inicialmente, los investigadores se centraron en problemas fundamentales como la comprensión de imágenes, el reconocimiento de objetos y el análisis de formas. A medida que evolucionó la visión por computadora, la estimación de pose surgió como un área de investigación distinta.

Antes de la llegada de las redes neuronales, la mayor parte del trabajo anterior se basa principalmente en estructuras pictóricas que modelan el cuerpo humano, como una colección de plantillas rígidas y un conjunto de potenciales por pares que toman la forma de una estructura de árbol [4].

Si bien los métodos tradicionales de visión por computadora proporcionaron información valiosa y allanaron el camino para avances posteriores, tenían limitaciones. A menudo eran sensibles al ruido, los valores atípicos, las oclusiones y dependían en gran medida de parámetros conocidos de la cámara. Como resultado, los enfoques basados en modelos comenzaron a ganar terreno. Los mismos utilizan modelos predefinidos de objetos o partes del cuerpo para estimar sus poses.

Más recientemente métodos basados en características se centraron en identificar y combinar características distintivas de la imagen para estimar la pose de objetos o sujetos humanos. Aprovecharon descriptores de características avanzados y técnicas de comparación para lograr una estimación de pose precisa y sólida.

Algunos ejemplos de estos algoritmos pueden ser Speeded Up Robust Features (SURF) y Scale-Invariant Feature Transform (SIFT) [5].

1.3.2. Aplicaciones

La estimación de Pose captó el interés de los investigadores debido a su amplia gama de aplicaciones en diversos dominios. Por ejemplo, en la interacción persona-computadora que permite a las computadoras interpretar y responder a los gestos humanos. Esto habilita una interacción intuitiva y natural entre humanos y máquinas para la detección de gestos o la animación de personajes en videojuegos. También encontró un nicho de uso en deportes y fitness, como vamos a ver en el trabajo, ayudando a analizar los movimientos corporales y la postura para mejorar el rendimiento y prevenir lesiones.

Algunas otras aplicaciones más específicas, pero no menos importantes, puede ser la experiencia de juegos interactivos, la realidad aumentada, el seguimiento de tiendas sin cajeros o en el monitoreo de actitudes agresivas [6][7].



FIGURA 1.3. Detección de peleas en la calle.

1.4. Estado del arte

Los métodos actuales de redes neuronales convolucionales (CNN por sus siglas en inglés), han demostrado una mejora sustancial en la solución de la gran variedad de problemas en visión por computadora. Superando ampliamente los modelos antes mencionados. Un aspecto clave de estos enfoques, es que integran la extracción de características jerárquicas no lineales, con la tarea de clasificación o regresión en cuestión. Pudiendo también capitalizar conjuntos de datos muy grandes que actualmente están disponibles. En el contexto de la estimación de la pose humana, es natural formular el problema como una regresión, en el que las características de CNN proporcionan una predicción conjunta de las partes del cuerpo [8].

Uno de los hitos clave en el campo de la estimación de la postura humana se produjo con la introducción de modelos como "PoseNet" desarrollado por Google [9]. El mismo se centró en abordar la relocalización de cámaras utilizando CNN,

pudiéndose adaptar el modelo para la estimación de posturas. Como así también la arquitectura "Stacked Hourglass" [10] específicamente diseñada para la estimación de poses.

Todos estos modelos se convirtieron en puntos de referencia en el campo. Otro avance significativo fue "OpenPos" [11], que emplea una arquitectura de red neuronal profunda para estimar la postura en tiempo real utilizando cámaras.



FIGURA 1.4. OpenPose estimación de pose 2D.

En los últimos años se desarrollaron varios frameworks de detección de postura, entre los que se destacan YOLOv7 [12] y MediaPipe [13] desarrollado por Google. El primero es un modelo de detección de objetos en tiempo real que se caracteriza por su capacidad para identificar múltiples objetos en una imagen o video, con una sola pasada. Su enfoque único de una sola etapa lo hace eficiente y rápido, lo que lo convierte en una excelente opción para aplicaciones en tiempo real. Por otro lado, MediaPipe se centra en la estimación precisa de la postura humana, utilizando una variedad de modelos para rastrear y detectar puntos clave en el cuerpo humano. Su enfoque en la estimación de posturas lo hace ideal para aplicaciones como reconocimiento de gestos y seguimiento de movimientos.

MediaPipe Pose es un marco de estimación de postura para una sola persona. Utiliza una topología de 33 puntos clave llamada BlazePose como puede verse en la figura 1.5. BlazePose es una topología que incluye puntos clave tanto de COCO como de Blaze Palm y Blaze Face, lo que la convierte en un conjunto más completo. El proceso de estimación de postura en MediaPipe Pose consta de dos etapas: detección y seguimiento:

- Detección: en esta etapa, el sistema detecta la presencia de una persona en la imagen y estima la posición inicial de los puntos clave.
- Seguimiento: una vez que se ha realizado la detección, el sistema realiza un seguimiento de los puntos clave en los frames sucesivos del video, lo que permite un seguimiento continuo de la postura.

Ambos frameworks tienen sus propias características y fortalezas, lo que los hace adecuados para diferentes aplicaciones y escenarios en el campo de la visión por computadora. Comparaciones realizadas en la práctica de yoga muestran que

YOLOv7 genera detecciones temblorosas como se ve en la figura 1.6 siendo además MediaPipe más eficiente en inferencias por CPU y debido a que sus detecciones no se realizan en todos los frames. Por lo tanto, lo hace la opción más atractiva para las aplicaciones en dispositivos de bajo poder de cómputo [14]. Estas son las razones por las que se eligió a MediaPipe para este proyecto.

Es importante destacar para el estado del arte que la estimación de la postura humana con Deep Learning sigue siendo un campo activo de investigación. En el cual constantemente surgen nuevas técnicas y enfoques para mejorar la precisión y la robustez de los modelos. Investigadores de todo el mundo contribuyen continuamente, lo que hace que sea un área en permanente evolución.

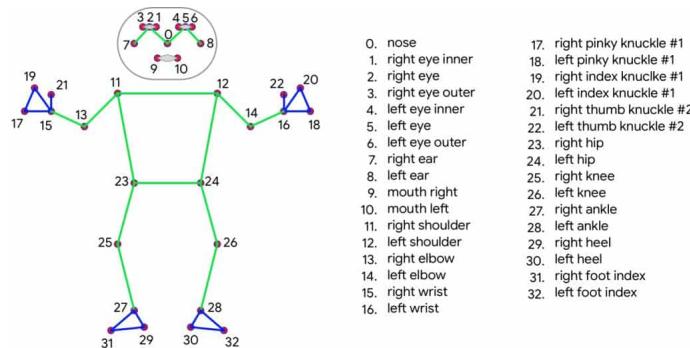


FIGURA 1.5. BlazePose topología de 33 puntos.

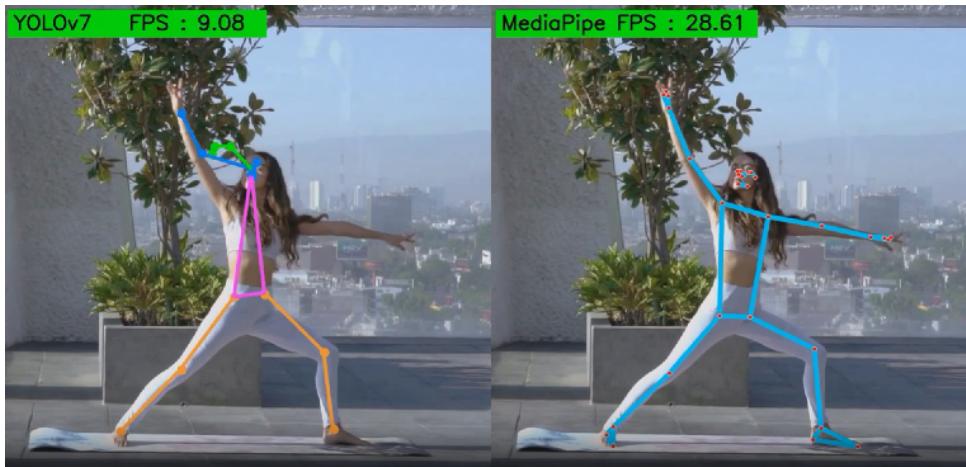


FIGURA 1.6. OpenPose estimación de pose 2D.

Capítulo 2

Introducción específica

En el presente capítulo se desarrolla una breve introducción a la tecnología de visión por computadora e inteligencia artificial. En particular, se verá la forma de funcionamiento del framework MediaPipe que se utilizó como modelo principal para la resolución del proyecto.

2.1. Técnicas de visión por computadora

Uno de los objetivos principales de la visión por computadora es la comprensión de escenas visuales. La comprensión de escenas implica numerosas tareas, que incluye el reconocimiento de qué objetos están presentes, su localización en 2D y 3D, la determinación de sus atributos y de la escena, la caracterización de las relaciones entre dichos objetos y la provisión de una descripción semántica de la escena [15]. La capacidad para extraer información valiosa de imágenes y videos convierte a este campo en una herramienta fundamental en la era digital actual.

A continuación, se presentan los conceptos básicos de visión por computadora y sus principales características:

- Adquisición de imágenes: la visión por computadora comienza con la obtención de datos visuales a través de cámaras digitales u otros dispositivos de captura de imágenes. Estos dispositivos convierten la luz visible en datos digitales que pueden ser procesados luego.
- Preprocesamiento de imágenes: antes de que las imágenes se puedan analizar, a menudo es necesario realizar una serie de pasos de preprocesamiento, como la corrección de la exposición, la eliminación de ruido y la normalización de colores. Esto garantiza que los datos sean coherentes y adecuados para su procesamiento.
- Extracción de características: la extracción de características implica identificar patrones o características relevantes como bordes, texturas, colores, formas o puntos de interés. Estas características sirven como entradas para algoritmos de análisis posteriores.
- Segmentación de imágenes: la segmentación se refiere a la tarea de dividir una imagen en regiones u objetos más pequeños. Esto puede ser útil para aislar objetos de interés o eliminar el fondo de una imagen.
- Reconocimiento de patrones: uno de los objetivos fundamentales de la visión por computadora es el reconocimiento de patrones, que implica la identificación y clasificación de objetos o elementos en una imagen. Esto se

logra mediante técnicas de aprendizaje automático, como redes neuronales convolucionales (CNN) o algoritmos de clasificación.

- Detección de objetos: la detección de objetos va más allá del reconocimiento y permite localizar la posición exacta de éstos en una imagen. Esto es crucial en aplicaciones como la conducción autónoma, la vigilancia y el seguimiento.
- Reconstrucción 3D: algunas aplicaciones de visión por computadora buscan reconstruir objetos tridimensionales a partir de imágenes bidimensionales. Esto se utiliza en campos como la realidad aumentada, la inspección industrial y la medicina.
- Seguimiento de movimiento: la visión por computadora también se utiliza para rastrear el movimiento de objetos o personas a lo largo del tiempo en secuencias de video. Esto es esencial en aplicaciones como el seguimiento de objetos en tiempo real y la vigilancia

2.2. Redes neuronales con capas convolucionales.

Las Redes Convolucionales se inspiraron en la estructura de la corteza visual, más específicamente, en el modelo del sistema visual descrito por Hubel y Wiesel en 1962 [16]. Los primeros modelos computacionales se basaron en tales interconexiones locales entre neuronas y transformaciones de imágenes estructuradas jerárquicamente.

Una CNN consta de tres tipos principales de capas neuronales: capas de convolución, capas de agrupación (pooling) y capas completamente conectadas (fully connected). Cada capa de una CNN traduce el volumen de entrada en un volumen de actividad neuronal de salida, lo que lleva a que los datos de entrada se traduzcan a la salida en un vector unidimensional. Se puede ver en la figura 2.1 la estructura básica de una CNN.

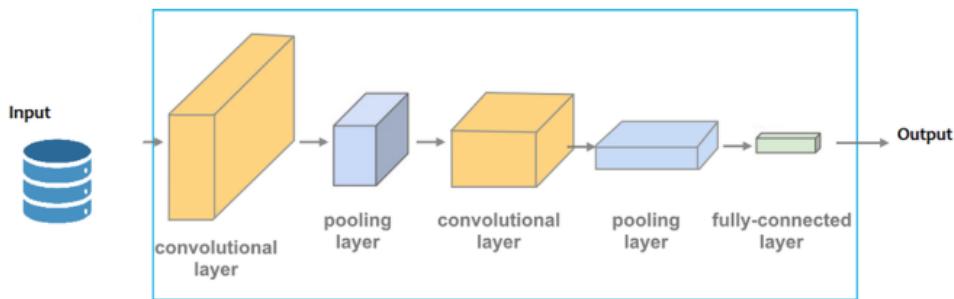


FIGURA 2.1. Estructura básica de una CNN.

A diferencia con otros modelos de redes neuronales, como el perceptrón de múltiples capas ocultas, la CNN está diseñada para aceptar múltiples matrices como entrada y procesarlas mediante operaciones de convolución. Debido a esto, sobresale en la resolución de los desafíos de visión por computadora, como la categorización, identificación y comprensión de imágenes. Para poder lograr alto rendimiento en la predicción y objetivos más exigentes la arquitectura de las CNN se vuelve más profunda y sofisticada [17].

Para explicar el comportamiento de una capa convolucional, aún necesitamos especificar algunos parámetros adicionales. Estos incluyen [18]:

- Padding: las redes tempranas como LeNet-5 no agregaban relleno a la imagen, por lo que ésta se reducía en tamaño después de cada convolución. Las redes modernas pueden especificar opcionalmente un ancho de relleno y un modo, como relleno de ceros o replicación de píxeles tal como se muestra en la figura 2.2.

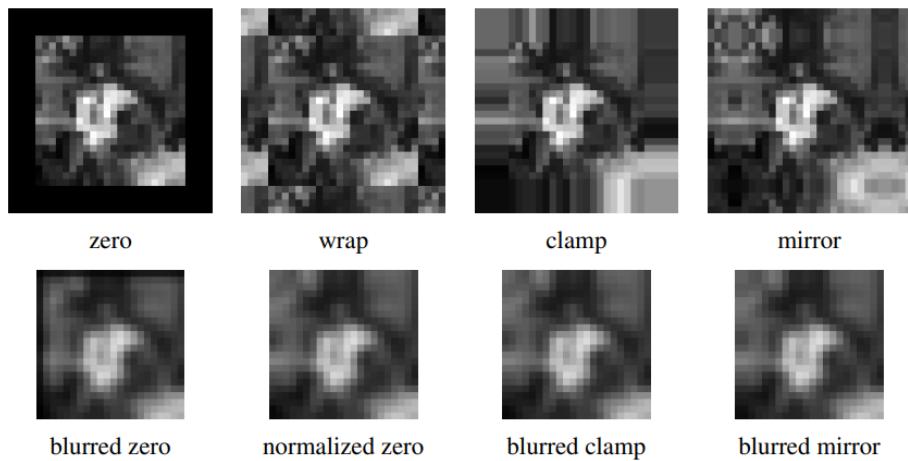


FIGURA 2.2. Relleno de borde (fila superior) y resultados de desenfocar la imagen rellenada (fila inferior). La imagen normalizada cero es el resultado de dividir la imagen RGBA borrosa con relleno de ceros por su correspondiente valor alfa suave.

- Stride: el paso predeterminado para la convolución es de 1 píxel, pero también es posible evaluar la convolución por cada columna y fila n-ésima. Por ejemplo, la primera capa de convolución de la de red AlexNet usa stride 4 como se ve en la figura 2.3. De esta manera al usar pasos superiores a uno es posible reducir la resolución de una capa determinada.

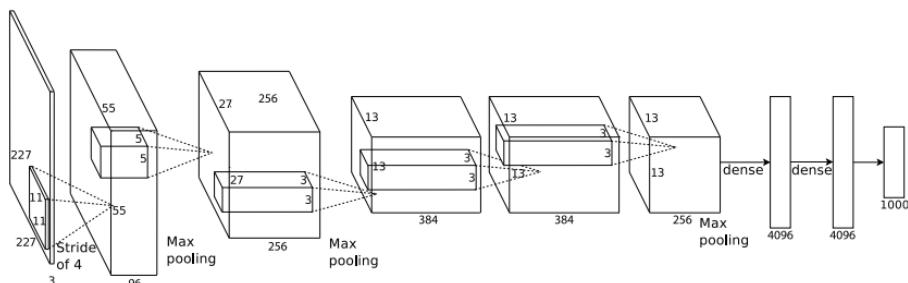


FIGURA 2.3. Arquitectura de la red neuronal profunda SuperVision (más comúnmente conocida como “AlexNet”). La red consta de múltiples capas convolucionales con activaciones ReLU, agrupación máxima, algunas capas completamente conectadas y una softmax para producir las probabilidades de clase finales.

- Dilatación: se puede insertar “espacios” adicionales (filas y columnas omitidas) entre las muestras de píxeles durante la convolución, también conocida como convolución dilatada. La dilatación puede ser efectiva para agrupar

una región más grande mientras se utilizan menos operaciones y parámetros aprendibles.

- Agrupamiento (*pooling*): por defecto, se utilizan todos los canales de entrada para producir cada canal de salida, pero también podemos agrupar las capas de entrada y salida en G grupos, cada uno de los cuales se convoluciona por separado. $G = 1$ corresponde a una convolución regular, mientras que $G = C_1$ significa que cada canal de entrada correspondiente se convoluciona de forma independiente de los demás, lo que se conoce como convolución *depthwise* o convolución separada por canales. El *pooling* y *unpooling* son operaciones esenciales en la construcción de redes CNN, ya que ayudan a reducir la dimensionalidad de las representaciones de imágenes y a preservar las características más importantes:

- *Pooling*: se utiliza para reducir la resolución de una capa dada al tomar una ventana cuadrada (como 2×2 o 3×3) y calcular el máximo o promedio dentro de esa ventana. El *max pooling* toma el valor máximo en la ventana, mientras que el *average pooling* toma el promedio. Estas capas de *pooling* ayudan a lograr invarianza espacial y reducen la dimensionalidad de la representación de la imagen.
- *Unpooling*: es el proceso inverso del *pooling*. Se utiliza para aumentar la resolución de una capa al agregar filas y columnas de ceros entre las muestras de píxeles en la capa de entrada. Luego, se aplica una convolución regular para aumentar el tamaño de la imagen. Esto se utiliza a menudo en tareas como la segmentación semántica para obtener una salida de mayor resolución.

2.3. Detección de objetos

La detección de objetos es el proceso de identificar la instancia de la clase a la que pertenece y establecer su ubicación, mostrando generalmente un cuadro delimitador a su alrededor. La detección puede ser de una o múltiples clases en donde el proceso identifica varios o todos los objetos de la imagen. Se puede clasificar ampliamente en dos categorías, detección de objetos dedicados en donde se encuentran la detección de rostros (figura 2.4), tráfico, peatones, etc. y por otro lado la detección de objetos genéricos.

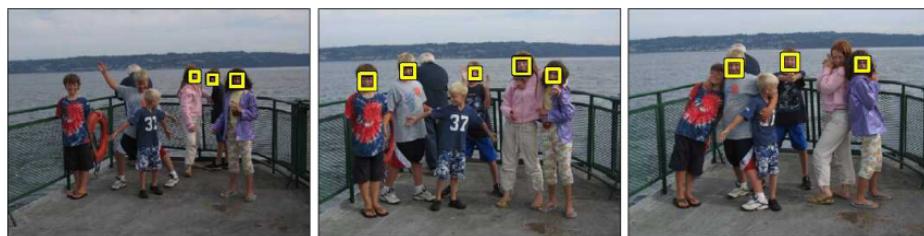


FIGURA 2.4. Moldelo de detección de rostro (Sivic, Zitnick, and Szeliski 2006).

El reconocimiento general de objetos y la comprensión de escenas plantea el desafío de realizar simultáneamente el reconocimiento y la segmentación con límites precisos. Mientras que la detección de objetos se basa en un conjunto de píxeles generales, la segmentación semántica utiliza los píxeles individuales asociados a

un objeto dado. Esto elimina el uso de cuadros delimitadores y permite una caracterización más precisa de los objetos en la imagen. La segmentación semántica suele hacer uso de redes totalmente convolucionales (FCNs) o redes neuronales no supervisadas (U-Nets).

El entrenamiento de vehículos autónomos es un uso notable para la segmentación semántica. Con este método, los investigadores pueden utilizar fotos de calles o carreteras con límites de objetos claramente definidos como se puede ver en la figura 2.5. Teniendo a la segmentación de instancias como la tarea de encontrar todos los objetos relevantes en una imagen y producir máscaras con precisión de píxel para sus regiones visibles.

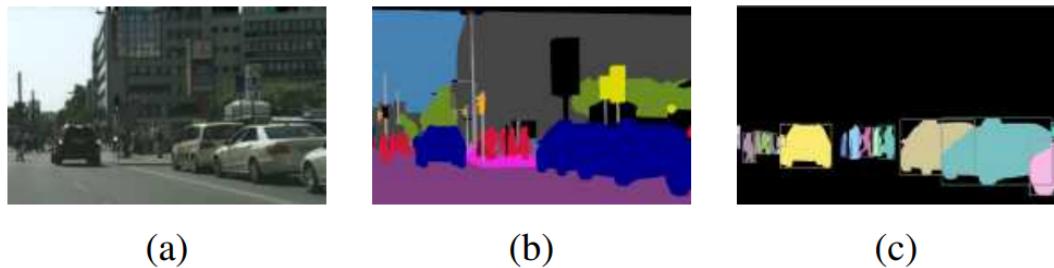


FIGURA 2.5. Ejemplos de segmentación de imágenes (a) imagen original; (b) segmentación semántica (clasificación por píxel); (c) segmentación de instancias (delinear cada objeto)

En lo que a este trabajo corresponde la inferencia de la pose humana (ubicación y actitud de la cabeza, el cuerpo y las extremidades) a partir de una sola imagen puede verse como otro tipo de tarea de segmentación. Cuando se fotografía a una persona u objeto, se utiliza una estimación de la pose para determinar dónde están las articulaciones y qué significa su postura. Funciona tanto con imágenes 2D como 3D. La arquitectura más utilizada para la estimación de la pose es PoseNet, un diseño basado en CNN [19] [20].

2.4. Modelado y seguimiento del cuerpo humano.

La modelización y el seguimiento del cuerpo humano, así como el reconocimiento de sus actividades, son algunas de las áreas más activamente estudiadas en visión por computadora. Dada la amplitud de esta área, resulta difícil categorizar toda esta investigación, especialmente debido a que las diferentes técnicas suelen construirse unas sobre la base de otras. Moeslund, Hilton y Kruger [21] dividen su estudio en inicialización, seguimiento (que incluye modelado de fondo y segmentación), estimación de postura y reconocimiento de acciones (actividades). Forsyth, Arikán y otros [22] dividen su investigación en secciones sobre seguimiento, sustracción de fondo, plantillas deformables, flujo y modelos probabilísticos. En base a esto se presentó una breve descripción de los temas mencionados [23]:

- Sustracción de fondo: uno de los primeros pasos en muchos sistemas de seguimiento de seres humanos consiste en modelar el fondo, para extraer los objetos en movimiento en primer plano.
- Inicialización y detección: para seguir a las personas de manera completamente automatizada, es necesario primero detectar (o volver a adquirir)

su presencia en los cuadros individuales de vídeo. Este tema está estrechamente relacionado con la detección de peatones, que a menudo se considera como una forma de reconocimiento de objetos.

- Seguimiento con flujo: el seguimiento de las personas y sus posturas en cada cuadro puede mejorarse mediante el cálculo del flujo óptico o la coincidencia de la apariencia de sus extremidades de un cuadro a otro. Por ejemplo, el modelo de personas de cartón de Ju, Black y Yacoob [24] modela la apariencia de cada porción de la pierna (superior e inferior) como un rectángulo en movimiento y utiliza el flujo óptico para estimar su ubicación en las imágenes subsiguientes.
- Modelos cinemáticos en 3D: la eficacia de la modelización y el seguimiento de seres humanos puede mejorarse significativamente mediante el uso de un modelo en 3D más preciso. Subyacente a tales representaciones, se encuentra un modelo cinemático o cadena cinemática, que especifica la longitud de cada miembro en un esqueleto, así como los ángulos de rotación entre los miembros o segmentos, figura 2.6(a-b). Se llama cinemática inversa a inferir los valores de los ángulos de las articulaciones a partir de las ubicaciones de los puntos superficiales visibles.

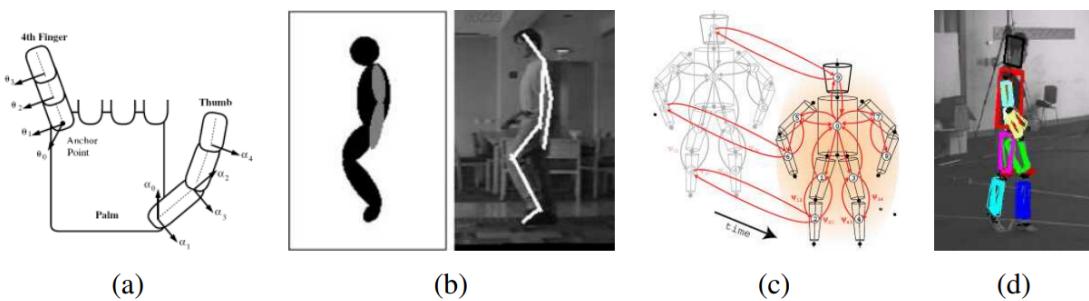


FIGURA 2.6. Seguimiento del movimiento humano en 3D: (a) modelo de cadena cinemática para una mano humana (b) seguimiento de un modelo de masa de cadena cinemática en una secuencia de video (c - d) colección probabilística de partes del cuerpo con miembros sueltos.

- Modelos probabilísticos: debido a que el seguimiento puede ser una tarea tan difícil, a menudo se utilizan sofisticadas técnicas de inferencia probabilística, para estimar los estados de la persona que se está siguiendo. Un enfoque popular, llamado filtrado de partículas [25], que fue desarrollado originalmente para el seguimiento de los contornos de personas y manos y posteriormente se aplicó al seguimiento de todo el cuerpo puede verse en la figura 2.6(c-d).
- Reconocimiento de actividad: el último tema ampliamente estudiado en la modelización humana es el reconocimiento de movimientos, actividades y acciones. Ejemplos de acciones comúnmente reconocidas incluyen caminar, correr, saltar, bailar, recoger objetos, sentarse, levantarse y saludar.

2.5. Descripción del modelo MediaPipe (BlazePoze)

MediaPipe es un framework específicamente diseñado para la estimación de postura de una sola persona que utiliza una serie de modelos. El primer modelo detecta la presencia de cuerpos humanos en un cuadro de imagen, y el segundo localiza puntos clave en los cuerpos. El framework opera en dos etapas: detección y seguimiento. La etapa de detección no se realiza en cada cuadro, lo que permite que el marco realice la inferencia de manera más rápida y eficiente.

2.5.1. Arquitectura del modelo y diseño del pipeline

A continuación se enumeran las principales características de la arquitectura del modelo extraídas del artículo [13]:

1. Inferencia: durante la inferencia, se emplea una configuración de detector-seguimiento o *tracker* que consta de un detector de postura corporal liviano previo de una red de seguimiento de postura como se muestra en diagrama de la figura 2.7. El *tracker* predice en el cuadro actual las coordenadas de los puntos clave, la presencia de la persona y la región de interés refinada. Cuando el *tracker* indica que no hay ninguna persona presente se vuelve a ejecutar la red de detección en el siguiente cuadro.

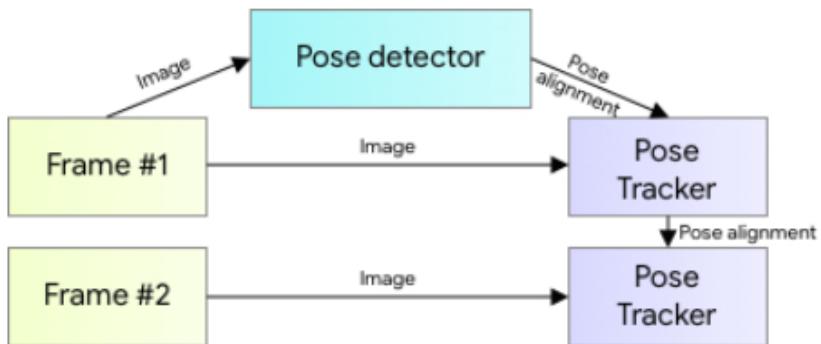


FIGURA 2.7. Pipeline de inferencia.

2. Detector de personas: la mayoría de las soluciones modernas de detección de objetos dependen del algoritmo de Supresión de No Máximo. Esto funciona bien para objetos rígidos con pocos grados de libertad. Sin embargo, este algoritmo se descompone en escenarios que incluyen poses altamente articuladas, por ejemplo, personas saludando o abrazándose. Esto se debe a que múltiples cajas ambiguas cumplen con el umbral de intersección sobre unión. Para superar esta limitación, el modelo se enfoca en detectar el rostro de la persona ya que se observó que en muchos casos es una característica fuerte que facilita la detección a la red neuronal. Es por esto que, haciendo la suposición sólida y válida de que la cara de la persona siempre debe ser visible, se utiliza un detector de rostros rápido en el dispositivo como sustituto de un detector de personas. Este detector de rostros predice parámetros adicionales de alineación específicos para la persona. Inspirado en el "Hombre de Vitruvio" de Leonardo da Vinci, determina el punto medio de las caderas, el radio del círculo que la circunscribe y el ángulo de inclinación de la línea que conecta los puntos medios de los hombros y las caderas. Esto resulta en un seguimiento consistente incluso en casos muy

complicados, como posturas específicas de yoga y se representa en la figura 2.8.

3. Topología: El modelo presenta una topología que utiliza 33 puntos del cuerpo humano. A diferencia de las topologías de OpenPose [26] y Kinect [27], utiliza solo el numero mínimo y suficiente de puntos clave en el rostro, las manos y los pies para estimar la rotación, el tamaño y la posición de la región de interés para el modelo posterior, como se puede ver en la figura 1.5.

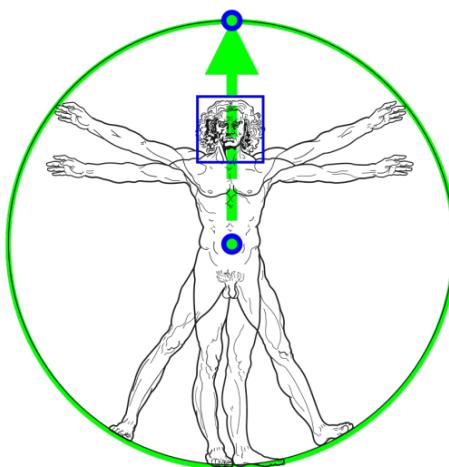


FIGURA 2.8. Hombre de Vitruvio alineado a través del detector versus cuadro delimitador de detección de rostros.

4. Dataset: En comparación con la mayoría de las soluciones de estimación de postura existentes, que detectan puntos clave utilizando mapas de calor, este modelo requiere una alineación inicial de la postura, es por esto que el conjunto de datos se limitó a casos en los que toda la persona es visible o donde los puntos clave de las caderas y los hombros pueden ser anotados con confianza. Para asegurar de que el modelo admite fuertes oclusiones, fue utilizada una amplia gama de técnicas de aumento de datos que simulan la oclusión sustancial. El conjunto de datos de entrenamiento constó de 60,000 imágenes con una o pocas personas en la escena en poses comunes y 25,000 imágenes con una sola persona en la escena realizando ejercicios. Todas estas imágenes fueron anotadas por seres humanos.
5. Arquitectura de red neuronal: El modelo adopta un enfoque combinado de mapas de calor (*heatmap*), desplazamientos (*offset*) y regresión, como se muestra en la figura 2.9. Las pérdidas de *heatmap* y *offset* se utilizaron solo en la etapa de entrenamiento y se remueven las correspondientes capas de salida del modelo antes de ejecutar la inferencia. De esta manera, se utilizan los *heatmap* de una forma efectiva para supervisar los *embedding* ligeros, que son luego utilizados por la red decodificadora de regresión.

Ademas la red utiliza conexiones de salto (*skip-connections*) activamente entre todas las etapas de la red para lograr un equilibrio entre características de alto y bajo nivel. Sin embargo, los gradientes de la red decodificadora de regresión no se propagan de vuelta a las características entrenadas en los *heatmap* (se puede observar las *gradient-stopping connections* de gradientes en la figura 2.9). Esto no solo mejora las predicciones de los *heatmap*,

sino que también aumenta sustancialmente la precisión de la regresión de coordenadas.

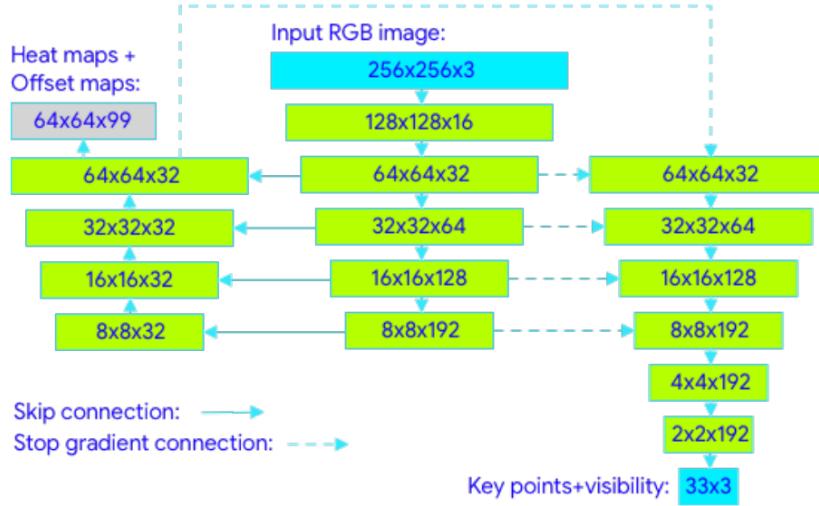


FIGURA 2.9. Arquitectura de red neuronal.

Este modelo de estimación de postura para una sola persona, fue desarrollado específicamente con el objetivo de utilizarse en casos que requieren un alto rendimiento como el lenguaje de señas, el seguimiento de posturas de yoga o prácticas de ejercicios físicos y también la realidad aumentada. Funciona en tiempo casi real en una CPU móvil y puede acelerarse para lograr una latencia mejorada a tiempo real en una GPU móvil.

Bibliografía

- [1] Melany Gabriela Estrella Pérez y Angie Anahy Peñafiel Vaca. «Análisis del impacto potencial del yoga en la salud física y mental como un descanso activo en los trabajadores de oficina en casa durante la pandemia de COVID-19». En: *UNIVERSIDAD SAN FRANCISCO DE QUITO - Colegio de Ciencias e Ingeniería* (2020).
- [2] Laura María Elena Miranda Hernández y Carmen Patricia Jiménez Terrazas. «La práctica del Yoga en el COVID-19: cambios y transformaciones». En: *REVISTA DOXA DIGITAL* (2021).
- [3] Adrian Bulat y Georgios Tzimiropoulos. *Human pose estimation via Convolutional Part Heatmap Regression*. Computer Vision Laboratory, University of Nottingham, 2016.
- [4] P.F. Felzenszwalb y D.P. Huttenlocher. «Pictorial structures for object recognition.» En: (2005).
- [5] Trevor Lynn. «Pose Estimation Algorithms: History and Evolution». En: *Roboflow* (2023).
- [6] Rafał Pytel. «Human Pose Estimation - 2023 guide». En: *Reasonfieldlab* (2023).
- [7] Mickael Cormier; Aris Clepe; Andreas Specker; Jürgen Beyerer. «Where are we with Human Pose Estimation in Real-World Surveillance?» En: *Winter Conference on Applications of Computer Vision Workshops (WACVW)* (2022).
- [8] Toshev, A. and Szegedy, C. «Deeppose: Human pose estimation via deep neural networks.» En: *CVPR* (2014).
- [9] Alex Kendall; Matthew Grimes and Roberto Cipolla. «PoseNet: A Convolutional Network for Real-Time 6-DOF Camera Relocalization». En: *University of Cambridge* (2015).
- [10] Alejandro Newell, Kaiyu Yang and Jia Deng. «Stacked Hourglass Networks for Human Pose Estimation». En: *TPAMI* (2019).
- [11] Zhe Cao, Student Member, IEEE, Gines Hidalgo, Student Member, IEEE, Tomas Simon, Shih-En Wei, and Yaser Sheikh. «OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields». En: *University of Cambridge* (2015).
- [12] Debapriya Maji, Soyeb Nagori, Manu Mathew, Deepak Poddar. «YOLO-Pose: Enhancing YOLO for Multi Person Pose Estimation Using Object Keypoint Similarity Loss». En: *Texas Instruments Inc* (2022).
- [13] Valentin Bazearevsky, Ivan Grishchenko, Karthik Raveendran, Tyler Zhu, Fan Zhang, Matthias Grundmann. «BlazePose: On-device Real-time Body Pose tracking». En: *Google Research* (2020).
- [14] Kukil, Vikas Gupta. «YOLOv7 Pose vs MediaPipe in Human Pose Estimation». En: *LearnOpenCV* (2022).
- [15] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick,

- Piotr Dollár. «Microsoft COCO: Common Objects in Context». En: *Microsoft* (2015).
- [16] D. H. Hubel and T. N. Wiesel. «Receptive fields, binocular interaction and functional architecture in the cat's visual cortex». En: *The Journal of Physiology*, (1962).
- [17] A. Srinivasan. «Handbook of Research on Computer Vision and Image Processing in the Deep Learning Era». En: SASTRA University (Deemed), Indiar, 2023, págs. 62-65.
- [18] Richard Szeliski. «Computer Vision: Algorithms and Applications». En: <https://szeliski.org/Book>, 2021, págs. 291-299.
- [19] A. Srinivasan. «Handbook of Research on Computer Vision and Image Processing in the Deep Learning Era». En: SASTRA University (Deemed), Indiar, 2023, págs. 93-94.
- [20] Richard Szeliski. «Computer Vision: Algorithms and Applications». En: <https://szeliski.org/Book>, 2021, págs. 387-396.
- [21] Hilton A. Moeslund T. B. y Kruger. «A survey of advances in vision-based human motion capture and analysis». En: *Computer Vision e Image Understanding*, 2006, 90-126.
- [22] Forsyth D. A. Arik O. Ikemoto L. O'Brien J. y Ramanan D. «Computational studies of human motion: Part 1, tracking and motion synthesis.» En: *Foundations, Trends in Computer Graphics y Computer Vision*, 2006, 77-254.
- [23] Richard Szeliski. «Computer Vision: Algorithms and Applications». En: <https://szeliski.org/Book>, 2021, págs. 843-850.
- [24] u S. X. Black M. J. y Yacoob Y. «Cardboard people a parameterized model of articulated image motion.» En: In *International Conference on Automatic Face y Gesture Recognition*, 1996, 38-44.
- [25] Isard M. y Blake A. «CONDENSATION conditional density propagation for visual tracking.» En: *International Journal of Computer Vision*, 1998, 5-28.
- [26] Z Cao, G Martinez Hidalgo, T Simon, SE Wei, and YA Sheikh. «Openpose: Realtime multi-person 2d pose estimation using part affinity fields.» En: *IEEE transactions on pattern analysis and machine intelligence* (2019).
- [27] Microsoft. *Azure kinect body tracking joints*. <https://docs.microsoft.com/en-us/azure/kinect-dk/body-joints>. Abr. de 2020.