EDUCACIÓN
SECRETARÍA DE EDUCACIÓN PÚBLICA

TECNOLÓGICO NACIONAL DE MÉXICO

Tecnológico Nacional de México

Centro Nacional de Investigación y Desarrollo Tecnológico

Tesis de Maestría

Estimación de la posición del cuerpo humano en secuencia de imágenes

presentada por

Ing. Luis Paul Arriaga Santos

como requisito para la obtención del grado de

Maestro en Ciencias de la Computación

Director de tesis

Dr. Raúl Pinto Elías

Cuernavaca, Morelos, México. Junio de 2024.

cenidet
Centro Nacional de Investigación y Desarrollo Tecnológico

Cuernavaca, Mor., 06/junio/2024

OFICIO No. DCC/069/2024
Asunto: Aceptación de documento de tesis
CENIDET-AC-004-M14-OFICIO

CARLOS MANUEL ASTORGA ZARAGOZA
SUBDIRECTOR ACADÉMICO
PRESENTE

Por este conducto, los integrantes de Comité Tutorial de **LUIS PAUL ARRIAGA SANTOS** con número de control M21CE052, de la Maestría en Ciencias de la Computación, le informamos que hemos revisado el trabajo de tesis de grado titulado **"ESTIMACIÓN DE LA POSICIÓN DEL CUERPO HUMANO EN SECUENCIAS DE IMÁGENES"** y hemos encontrado que se han atendido todas las observaciones que se le indicaron, por lo que hemos acordado aceptar el documento de tesis y le solicitamos la autorización de impresión definitiva.

Raúl Pinto Elías
Director de tesis

Andrea Magadán Salazar
Revisor 1

Nimrod González Franco
Revisor 2

C.c.p. Depto. Servicios Escolares.
Expediente / Estudiante

Cuernavaca, Mor., 11/junio/2024
No. De Oficio: SAC/189/2024
Asunto: Autorización de impresión de tesis

LUIS PAUL ARRIAGA SANTOS
CANDIDATO AL GRADO DE MAESTRO
EN CIENCIAS DE LA COMPUTACIÓN
P R E S E N T E

Por este conducto, tengo el agrado de comunicarle que el Comité Tutorial asignado a su trabajo de tesis titulado **"ESTIMACIÓN DE LA POSICIÓN DEL CUERPO HUMANO EN SECUENCIAS DE IMÁGENES"**, ha informado a esta Subdirección Académica, que están de acuerdo con el trabajo presentado. Por lo anterior, se le autoriza a que proceda con la impresión definitiva de su trabajo de tesis.

Esperando que el logro del mismo sea acorde con sus aspiraciones profesionales, reciba un cordial saludo.

A T E N T A M E N T E
Excelencia en Educación Tecnológica®
"Conocimiento y tecnología al servicio de México"

EDUCACIÓN | TECNOLÓGICO NACIONAL DE MÉXICO
CENTRO NACIONAL DE INVESTIGACIÓN Y DESARROLLO TECNOLÓGICO
SUBDIRECCIÓN ACADÉMICA

CARLOS MANUEL ASTORGA ZARAGOZA
SUBDIRECTOR ACADÉMICO

C. c. p. Departamento de Ciencias Computacionales
Departamento de Servicios Escolares

CMAZ/lmz

Dedicatorias

Este trabajo de investigación va dedicado a mi familia quien han estado incondicionalmente durante todo el proceso de mi desarrollo profesional

Con amor se lo dedico a mi madre, el principal apoyo emocional para seguir adelante con mi desarrollo personal y profesional.

A mi padre, quien, pese a las adversidades, nunca ha dejado de impulsarme a ser mejor persona.

A mi hermana Jennifer, por siempre ser un pilar para los éxitos alcanzados, y por siempre mostrar su apoyo para cumplir mis metas.

A Lizbeth por ser mi amiga y dulce compañera de aventuras.

Agradecimientos

Al Consejo Nacional de Humanidades Ciencia y Tecnología (CONAHCYT) por la beca otorgada para la realización de esta investigación.

Agradezco a todo el personal del TecNM / CENIDET por la hospitalidad durante mi estancia en este centro de investigación.

Quiero Agradecer a mi director de tesis, el Dr. Raúl Pinto Elías por su conocimiento y guía durante el proceso de investigación.

De igual manera quiero agradecer a mi comité revisor, la Dra. Andrea Magadán Salazar y al Dr. Nimrod González Franco por las observaciones y sugerencias para la presentación final de este trabajo.

Finalmente, quiero agradecer a mi familia por apoyarme para cumplir esta meta en mi vida profesional.

Resumen

La estimación de la pose humana brinda una rama de exploración distinta a las utilizadas para el reconocimiento de actividades humanas (**HAR**, por sus siglas en inglés), los enfoques principalmente utilizados son basados en el uso de sensores con placas de desarrollo, dispositivos inteligentes o a través del procesamiento de imágenes.

En este trabajo se propone implementar un modelo de estimación de la pose y otro de clasificación capaces de localizar y clasificar un conjunto de poses humanas a través de puntos característicos del cuerpo humano.

Para determinar el modelo de estimación de la pose humana (HPE) a utilizar se realizó la comparativa entre cuatro modelos de estimación de la pose disponibles en la literatura, *OpenPose, HRNet, YoloPose y BlazePose.* Esto con ayuda de Open MMPose para la evaluación de los modelos en dos conjuntos de datos CrowdPose y OcHuman.

La experimentación para seleccionar un modelo de clasificación de los puntos característicos se realizó en dos conjuntos de datos para la etapa de clasificación, uno de origen propio, así como un conjunto de datos reportado en la literatura; se evaluaron los algoritmos de *random forest, logistic regression, ridge classifier* y *gradient boosting classifier* con las métricas de exactitud, precisión y sensibilidad.

Los resultados muestran que la metodología seguida brinda excelentes resultados para la construcción de un sistema capaz de estimar y clasificar la pose huma aun con la presencia de oclusiones parciales.

Abstract

Human pose estimation provides a distinct branch of exploration compared to those used for Human Activity Recognition (HAR). The primarily used approaches are based on the use of sensors with development boards, smart devices, or through image processing.

In this work, the implementation of a pose estimation model and a classification model is proposed, capable of locating and classifying a set of human poses through key points of the human body. To determine the human pose estimation (HPE) model to be used, a comparison was conducted among four pose estimation models available in the literature: OpenPose, HRNet, YoloPose, and BlazePose. This was done with the assistance of Open MM Pose for model evaluation on two datasets, CrowdPose and OcHuman.

The experimentation for selecting a key point classification model was carried out on two datasets for the classification stage, one of our own origins and one reported in the literature. The algorithms of random forest, logistic regression, ridge classifier, and gradient boosting classifier were evaluated using metrics such as accuracy, precision, and sensitivity.

The results demonstrate that the followed methodology yields excellent results for the development of a system capable of estimating and classifying human pose, even in the presence of partial occlusions.

Índice

Lista de figuras

Lista de tablas

Glosario

Palabra	Definición
Top-Down	Enfoque Divisivo, localiza primero a las personas y posteriormente localiza los puntos característicos
Bottom-Up	Enfoque Aglomerativo, localiza todos los puntos característicos y posteriormente los agrupa en individuos
BlazePose	Modelo de estimación de la pose
Landmarks	Puntos clave o característicos, mayormente articulaciones
Frame	Cuadro de video
BlazeFace	Modelo de estimación de puntos característicos faciales
BlazePalm	Modelo de estimación de puntos característicos en manos
OpenPose	Modelo de estimación de la pose *bottom up*
YoloPose	Modelo de estimación de la pose basado en arquitectura de Yolo
HRNet	Modelo de estimación de la pose basado en la arquitectura High Resolution Network
Random Forest	Modelo de Aprendizaje automático basado en árboles aleatorios
Linear Regression	Modelo de Aprendizaje automático basado en regresión lineal
Ridge Classifier	Modelo de aprendizaje automático basado en clasificador de cresta
Gradient Boosting Classifier	Modelo de aprendizaje automático basado en una técnica de conjunto (aprendizaje ensamblado)
ROIPooling	Extracción de características relevantes de regiones específicas de una imagen, conocidas como "regiones de interés" (RoI)

Palabra	Definición
Faster-RCNN	Arquitectura de red neuronal convolucional (CNN) diseñada para la detección rápida y precisa de objetos en imágenes
Backbone	Parte principal o fundamental de la arquitectura de una red neuronal, extracción de características
***S*tacked hourglass**	arquitectura de red neuronal utilizada comúnmente en el campo de la visión por computadora
K-means	Modelo de aprendizaje automático no supervisado
Res*N*et50	Arquitectura de red convolucional residual con 50 capas
Transformer	Mecanismo de "atención", que permite procesar secuencias de datos de longitud variable
TransPose	Modelo de estimación de la Pose humana
Dataset	Conjunto de datos

Acrónimos

Acrónimo	Significado
RGB	Red, Green,Blue (Rojo, Verde, Azul)
2D	2 dimensiones
3D	3 dimensiones
HAR	Human Activity Recognition (Reconocimiento de Actividades humanas)
HPS	Human Pose Estimation (Estimación de la pose)
PAF	Part Afinity Fiels (Campos de Afinidad de Partes)
YoLo	You Only Look Once (Solo iras una vez)
VP	Verdadero Positivo
VN	Verdadero Negativo
FP	Falso Positivo
FN	Falso Negativo
OKS	Object Keypoint Similarity (Similitud de Puntos Clave de Objetos)
COCO	Common Objects in Context (Objetos comunes en contexto)
AP	Average Precision (Promedio de precisión)
ROI	Region Of Interest (Región de Interés)
NMS	No Maximun Supresion (Supresión de no Máximos)
LiDar	Light Detection and Ranging (Detección de Luz y Rango)
PCK	Percentage of Correct Keypoints

Acrónimo	Significado
DEKR	*DisEntangled Keypoint Regression (Regresión de Puntos Clave Desentrelazada)*
DNN	Deep Neural Network (Red Neuronal Profunda)
CPN	Convolutional Pose Machine (Máquina de Poses Convolucional)
IoU	Intersection over Union (Intersección sobre Unión)
mAP	mean Average Precision (Media de Precisión Promedio)
MPJPE	Mean Per Joint Position Error (Error Promedio por Posición de Articulación)
CNN	Convolutional Neural Network (Red Neuonal Convolucional)
GPU	Graphic Unit Process (Unidad de Rrocesamiento Gráfico)
CPU	Central Process Unit (Unidad de Procesamiento Central)
HRNet	High Resolution Network (Red de alta resolución)
PCP	Percentage of Correct Parts (Porcentaje de partes correctas)
PDJ	Percentage of Detected Joints (Porcentaje de unions detectadas)
PCK	Percentage of Correct Key-points (Porcentaje de puntos claves correctos)
PCKh	Percentage of Correct Key-points Head normalized (Porcentaje de puntos claves correctos normalizados)

CAPÍTULO 1
INTRODUCCIÓN

El reconocimiento de actividades humanas (HAR, por sus siglas en inglés) es un campo de estudio dentro de la inteligencia artificial y ciencias computacionales que se enfoca en identificar y clasificar diversas actividades humanas. HAR suele implicar sensores y algoritmos para identificar patrones en el comportamiento humano.

HAR se usa en muchas aplicaciones, como cuidado de la salud, sistemas de monitoreo de seguridad, análisis de comportamiento de individuos. Los enfoques usados en HAR varían según su aplicación propia de cada sistema, pero son comunes los giroscopios, acelerómetros y magnetómetros, incorporados mediante placas de desarrollo o dispositivos inteligentes, lo que implica una compleja caracterización de los sensores para disminuir posibles perturbaciones causadas por ruido en el ambiente.

Otro enfoque es el basado en el procesamiento de imágenes, ya sea la imagen completa o a través de la extracción de un subconjunto de puntos característicos a partir de la estimación de la pose humana. La estimación de la pose humana es un área bastante activa en la visión por computadora, que da grandes ventajas sobre otros enfoques utilizados para el reconocimiento de actividades, como la invariabilidad con respecto a rotaciones, cambios de iluminación, perspectivas y oclusiones.

Considerando lo mencionado, se realizó la clasificación de la pose humana a través de un conjunto de puntos característicos, así como la codificación de estos a través de medidas entre los puntos para implementar algoritmos de aprendizaje automático para la clasificación de la pose y de igual manera su evaluación con presencia de oclusiones artificiales parciales y autooclusiones.

1.1 Descripción del problema

Al localizar el cuerpo para estimar su pose en un escenario tridimensional, hay factores que provocan la oclusión de la persona a localizar, lo que trae una baja eficiencia en la estimación de la pose, ya que la estimación recae en la localización de las personas o articulaciones encontradas.

1.1.1 Delimitación del problema específico

El principal problema es estimar la pose del cuerpo humano aun con presencia de oclusiones parciales en la imagen, secuencia de imágenes o video; posteriormente clasificar los puntos característicos extraídos de la pose en un conjunto de poses ya definidas.

1.1.2 Complejidad del problema

Los siguientes puntos muestran la complejidad del problema:

- Cada imagen puede contener un número indeterminado de personas que pueden aparecer en cualquier posición o escala (Huang, Huang and Tang, 2021).
- La interacción entre personas induce a un espacio complejo de interferencias (Huang, Huang and Tang, 2021)
- La estimación de la pose humana en 3D a partir de una imagen RGB monocular es todavía una tarea retadora debido a la ambigüedad (Chen, Tian and He, 2020).
- Avances en la estimación de la posición humana en 3D aún se mantienen limitados debido a la auto oclusión, a una débil generalización y a una ambigüedad inherente de la recuperación de la profundidad (Kato, Honda and Uchida, 2020).
- La presencia de sujetos vistiendo ropa holgada, como: faldas largas, vestidos o capas, donde las extremidades están parcialmente o completamente cubiertas y no

son distinguibles, deja una oportunidad de trabajo para el reconocimiento de la posición humana (Dang *et al.,* 2019a).

1.2 Objetivos

Objetivo general

- Diseñar un sistema de visión artificial que estime un conjunto de poses, a pesar de presentar oclusiones parciales, a partir de una imagen o secuencia de imágenes.

Objetivos específicos

- Diseñar un sistema de visión artificial autónomo capaz de estimar la pose humana en un área controlada.
- Revisar en el estado del arte las técnicas utilizadas para la estimación de la pose.

1.3 Alcances y limitaciones

Alcances

- A través de vistas parciales el sistema será capaz de estimar la posición del cuerpo humano sin importar las oclusiones parciales presentes.
- Se identificarán de 5 a 10 poses que se presenten en la secuencia de imágenes.
- Las imágenes analizadas deberán contener al menos una persona.

Limitaciones

- El sistema únicamente detectará a personas que no presenten oclusiones totales.
- Dentro de un escenario con múltiples personas, únicamente se detectará a una persona a la vez.
- La estimación de la pose únicamente se dará en un ambiente controlado

1.4 Marco conceptual

En esta sección se presenta la parte teórica requerida para solucionar los objetivos planteados con la metodología mencionada, así como una explicación de los modelos de clasificación y estimación de la pose utilizados y evaluados en la etapa de experimentación en el capítulo 4 de este documento.

1.4.1 Estimación de la pose

La estimación de la pose humana es un problema de predicción de puntos característicos del cuerpo de una persona en una imagen; estos puntos característicos son asociados a articulaciones del cuerpo, tales como hombros, rodillas, muñecas y más, sin embargo, también se asocian a puntos relativos capaces de influenciar en la localización de dichas articulaciones (Dang *et al.,* 2019b).

Existen dos paradigmas principales para la estimación de la pose: *Top-down y Bottom-up*. El paradigma *top-down* localiza primero a la persona y posteriormente ejecuta la localización de los puntos característicos, mientras que el enfoque bottom-up localiza primero todos los puntos característicos visibles en la imagen y posteriormente los agrupa en individuos. El paradigma *top-down* es más exacto, pero con mayor costo computacional (Geng *et al.,* 2021a).

En la literatura existen varios modelos de estimación de la pose, entre ellos:

- BlazePose (Bazarevsky and Grishchenko, 2020a).

Esta una arquitectura basada en redes neuronales convolucionales para la estimación de la pose humana, adaptada para ejecutarse en tiempo real en dispositivos móviles.

Consiste de la estimación de puntos característicos a partir del proceso de inferencia plantado en la Figura 1. 1.

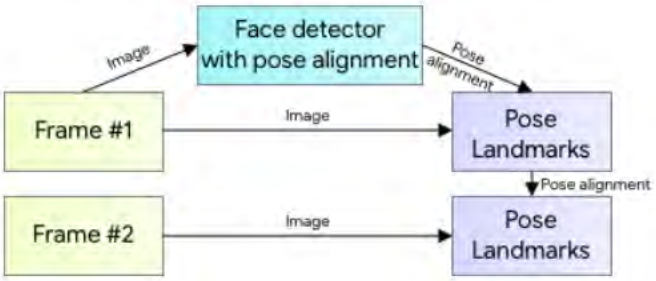

Figura 1. 1 Flujo de inferencia en BlazePose (Bazarevsky and Grishchenko, 2020a)

Este modelo parte de la premisa de la presencia del rostro para realizar la estimación de la pose, por lo que incorpora su modelo de *BlazeFace y BlazePalm* para la identificación de los 33 puntos característicos como se muestra en la Figura 1. 2.

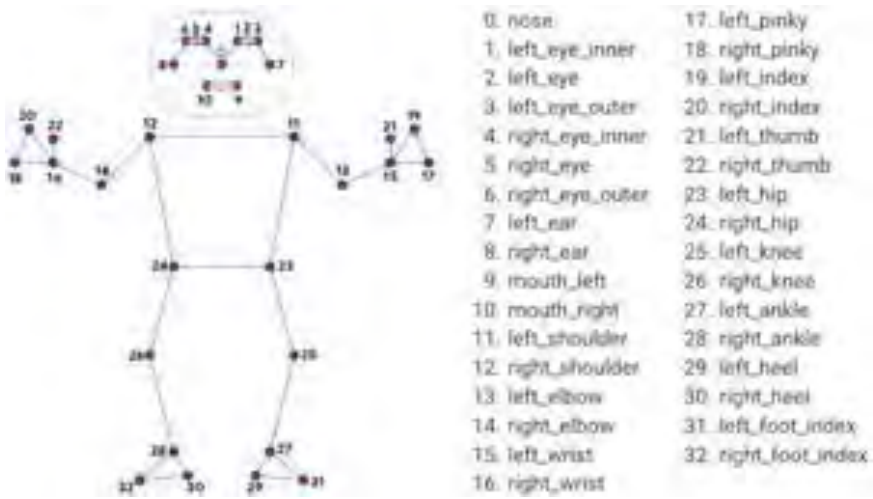

Figura 1. 2 Puntos característicos dados por BlazePose (Bazarevsky and Grishchenko, 2020)

- OpenPose (Cao *et al.*, 2021)

Este modelo de estimación de la pose humana 2D sigue un enfoque *bottom-up*, usando una representación no paramétrica referida como campos de afinidad de partes (PAFs,

por sus siglas en inglés) para aprender a asociar partes del cuerpo entre individuos. Esto mantiene una alta exactitud manteniendo un rendimiento en tiempo real.

El modelo realiza la estimación de 16 puntos característicos por individuo y muestra un tiempo de ejecución constante conforme aumenta el número de personas en la imagen, como se muestra en la Figura 1. 3. Más detalles de este modelo se pueden revisar en el trabajo original planteado en (Cao *et al.*, 2017a).

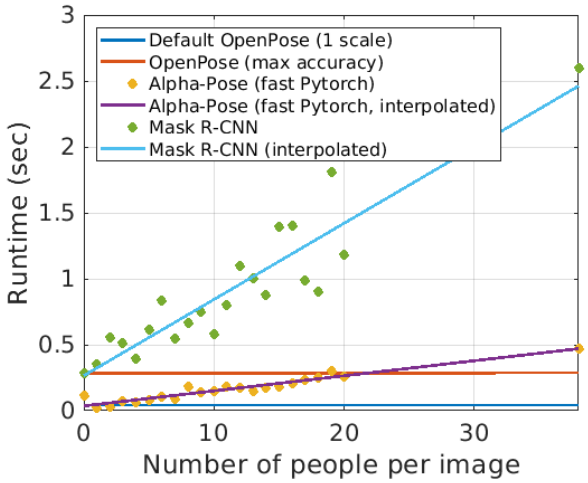

Figura 1. 3 tiempo de ejecución contra número de personas con OpenPose *(Cao* et al.*, 2021)*

- YoloPose (Debapriya M, Soyeb N and Manu M, 2022a)

Este modelo basado en la arquitectura de Yolo permite realizar la estimación de la pose humana en 2D sin el uso de mapas de calor como se plantea en muchos trabajos, haciendo esto en dos pasos, los cuales se asocian a la metodología *top-down* o *bottom-up*; siendo esto un factor que no permite el entrenamiento de los modelos de principio a fin sin etapas medias.

Este modelo realiza la detección de los cuadros delimitadores de las personas y su pose correspondiente en un solo proceso. La Figura 1. 4 muestra la detección de la persona y su estimación de la pose humana. Más detalles del modelo se pueden revisar en (Debapriya M, Soyeb N and Manu M, 2022b).

Figura 1. 4 Estimación de la pose humana realizada por el modelo de YoloPose (Debapriya M, Soyeb N and Manu M, 2022a)

- HRNet (Wang *et al.,* 2021)

Este modelo de estimación de la pose humana está basado en la arquitectura planteada en (Sun *et al.,* 2019), donde se usa el aprendizaje de una representación en alta resolución fiable para la estimación, esto manteniendo la alta resolución durante toda la imagen.

El modelo utiliza el enfoque *top-down* para la estimación de la pose siguiendo un proceso en paralelo para la identificación de los mapas de confianza, fue evaluado en el conjunto de datos COCO por lo que realiza la estimación de 17 puntos característicos.

1.4.2. Clasificación de la pose

La clasificación de la pose humana se realizó a través de cuatro algoritmos de aprendizaje automático, dos basados en regresión y dos en árboles de decisión, estos se describen brevemente a continuación.

- *Random Forest*

Random forest es un algoritmo de aprendizaje automático usado para la clasificación y regresión. Es una técnica de aprendizaje ensamblado que combina a múltiples árboles de decisión para realizar predicciones. Detalles sobre el algoritmo pueden ser revisados con (Breiman, 2001).

- *Linear Regression*

Linear Regression es un algoritmo de aprendizaje automático utilizado para la clasificación de variables categóricas o binarias. A diferencia de la regresión lineal, la regresión logística se utiliza para predecir la probabilidad de que una observación pertenezca a una de las clases posibles. Detalles sobre el algoritmo pueden ser revisados con (Sabouri *et al.,* 2020).

- *Ridge Classifier*

Ridge Classifier es un algoritmo de clasificación lineal utilizado en aprendizaje automático. Es una variante del algoritmo de regresión lineal que utiliza una técnica de regularización para evitar el sobreajuste en modelos lineales. Detalles sobre el algoritmo pueden ser revisados en (Deepa *et al.,* 2021).

- *Gradient Boosting Classifier*

Gradient Boosting Classifier es un algoritmo de aprendizaje automático utilizado en problemas de clasificación. Es un tipo de algoritmo ensamblado que combina múltiples modelos más simples, en este caso árboles de decisión, para crear un modelo de clasificación más preciso y robusto. Detalles sobre el algoritmo pueden ser revisados en (Alshari, Saleh and Odabaş, 2021).

1.4.3. Métricas

Para evaluar el desempeño de los modelos de clasificación se han utilizado tres métricas principales para el desarrollo de modelos basados en aprendizaje supervisado, como es el caso de este proyecto. Las métricas son: exactitud, precisión y sensibilidad, dichas métricas utilizan la matriz de confusión como la que se muestra en la Tabla 1.1 como base para su cálculo, donde la matriz de confusión es una representación visual tabular de las predicciones del modelo contra los valores reales de las etiquetas. Cada columna de la matriz de confusión representa las instancias de una clase predicha, mientras que cada fila representa las instancias de la clase real u observación (Bharadwaj, Prakash and Kanagachidambaresan, 2021).

Tabla 1. 1 Matriz de confusión

		Predicción	
		Positivos	**Negativos**
Observación	**Positivos**	Verdaderos Positivos (VP)	Falsos Negativos (FN)
	Negativos	Falsos Positivos (FP)	Verdaderos Negativos (VN)

Exactitud:

$$\frac{VP + VN}{(VP + FP + FN + VN)} \quad (1)$$

Precisión:

$$\frac{VP}{(VP + FP)} \quad (2)$$

Sensibilidad:

$$\frac{VP}{(VP + FN)} \quad (3)$$

Donde las literales se describen como:

- VP = Verdaderos positivos
- VN = Verdaderos negativos
- FP = Falsos positivos
- FN = Falsos negativos

Cabe mencionar que para la validación de los modelos de estimación de la pose se calculó la precisión media a partir de la métrica OKS (*Object Keypoint Similarity),* la cual es una

medida cuantitativa de la similitud entre los puntos característicos predichos y los realmente anotados en el conjunto de datos.

$$OKS = \frac{\exp\left(\frac{-d_i^2}{2s^2k_i^2}\right)\delta(v_i>0)}{\sum_i \delta(v_i>0)} \quad (4)$$

Donde las literales se describen como:

- d_i = distancia euclidiana entre cada punto correspondiente del valor predicho y el real.
- V_i = banderas de visibilidad en los puntos reales.
- $s\kappa_i$ = para el cálculo de OKS se utiliza una función gausina no nomalizada con una desviación estándar, deonde S es el objeto de scala y κ_i es una constante por punto clave que controla la caída

La métrica OKS considera la distancia entre los puntos característicos correspondientes y la escala de la persona que se evalúa. Calcula una puntuación entre 0 y 1, donde 0 representa ninguna similitud y 1 representa una coincidencia perfecta. Puntuaciones más altas de OKS indican una mayor precisión en la estimación de pose. La métrica OKS se utiliza a menudo en conjuntos de datos de referencia y desafíos, como COCO (*Common Objects in Context*), para evaluar y comparar el rendimiento de diferentes modelos y algoritmos de estimación de pose.

AP^{50} y AP^{75} miden específicamente la precisión media cuando la sensibilidad es al 50% y 75% respectivamente, esto quiere decir que solo los valores de precisión alcanzados al umbral de del 50% y 75% son considerados para el cálculo de la precisión media.

1.4.4. Distancia Manhattan

La distancia manhattan es utilizada en este trabajo para calcular la distancia entre pares de puntos característicos como se muestra en la Figura 1. 5. esta distancia se calcula

utilizando las coordenadas cartesianas de los puntos característicos con respecto a (*x,y*) con la finalidad de obtener una caracterización de los puntos clave que sea invariante a rotaciones y traslaciones.

Distancia Manhattan:

$$|A - B| = \Sigma_{i=1}^{d} |a_i - b_i| \quad (4)$$

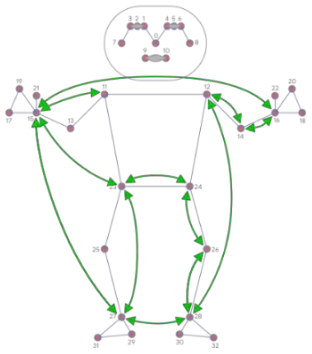

Figura 1. 5 Distancias calculadas entre puntos característicos

Donde la ecuación se describe como:

- $|A - B|$ = Distancia absoluta entre los puntos clave A y B
- $\Sigma_{i=1}^{d} |a_i - b_i|$ = La sumatoria de absoluta de cada punto independiente $a_i \; y \; b_i$ correspondientemente.

1.5 Organización de la tesis

En el capítulo 2 se presenta el estado del arte relacionado a las técnicas, modelos, aplicaciones y problemas de oclusión presentes en la estimación de la pose humana; el uso de la estimación de la pose humana para el reconocimiento de actividades humanas, los antecedentes dentro del CENIDET y discusión del estado del arte.

El capítulo 3 presenta el análisis teórico sobre el cual se ha sustentado la metodología para el desarrollo del proyecto, así como los recursos, bibliotecas y conjuntos de datos utilizados.

En el capítulo 4 se presenta la experimentación a través de la metodología planteada, así como los resultados obtenidos.

Por último, el capítulo 5 menciona las conclusiones obtenidas, así como el cumplimiento de los objetivos. De igual manera se detalla las aportaciones, actividades académicas y trabajos futuros como sugerencia del desarrollo de la investigación.

CAPÍTULO 2 ESTADO DEL ARTE

En este capítulo se presenta trabajos que han sido utilizados como referencia para el desarrollo de esta investigación, divididos en diferentes enfoques asociados a la identificación de técnicas, modelos y métricas para la estimación de la pose humana, así como aplicaciones y usos dentro del reconocimiento de actividades humanas a partir de la estimación de la pose.

El estudio se basó en la revisión de al menos 50 artículos asociados a la estimación de la pose en sus diferentes técnicas, generando así un documento con los 25 modelos encontrados dentro de la literatura para estimar la pose del cuerpo humano

2.1. Técnicas de Estimación de la pose

Dentro de la literatura existen trabajos asociados a modelos y técnicas utilizados para la estimación de la pose, entre ellos destacan los siguientes trabajos presentados en esta sección.

2.1.1 Research on 3D Human Pose Estimation Using RGBD Camera(Tang, Wang and Chen, 2019a)

En el artículo se plantea la implementación de una cámara RGB-D para la estimación de la pose a partir de una imagen a color y la información de profundidad. Como principal planteamiento se dice que la estimación de la pose a partir de una sola imagen a color no muestra resultados precisos debido a una falta de la información de la profundidad.

Plantean dos enfoques para estimar la pose 3D, el primero mencionado es a partir de una imagen a color con las anotaciones en dos dimensiones de las articulaciones encontradas y otra con múltiples cámaras en paralelo (Visión estéreo, por ejemplo).

La estimación 3D se realiza a partir de la estimación en 2D, por lo que la precisión de la estimación en 3D recae directamente en su precisión en 2D. Se utiliza un enfoque *top-down.*

Se decidió utilizar una estructura general de la *Faster-RCNN* para la detección de personas, la cual está formada por tres partes principales, la extracción de características, la estructura RPN y la estructura *ROIPooling*. En la Figura 2. 1 se muestra el modelo planteado con sus diferentes etapas para la estimación de la pose.

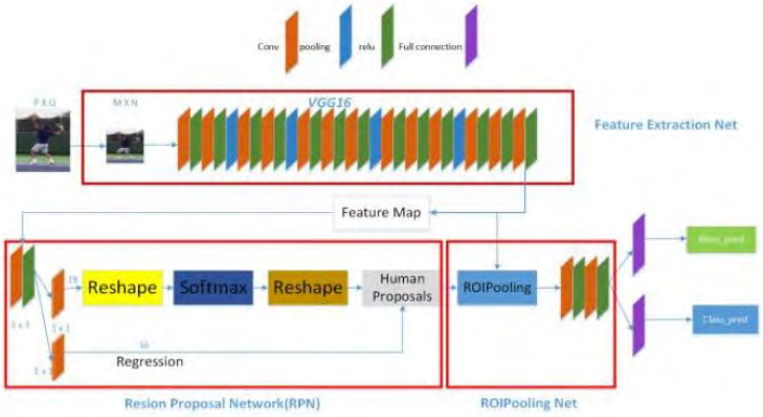

Figura 2. 1 Modelo planteado (Tang, Wang and Chen, 2019b)

Como red *backbone* se decidió utilizar *Resnet50* para la extracción de características. Para la tarea de la estimación de la pose, se utilizó una red neuronal convolucional *stacked hourglass.*

Se planteó un proceso de alineación de la imagen a color con la imagen de profundidad, por la diferencia en las resoluciones, como ejemplo en el Kinect v2 de Microsoft, donde la resolución de la imagen a color es de 1920*1080, y la resolución de la imagen de profundidad es de 512*424, provocando errores.

Para este proceso de alineación se propone un método de calibración basado en puntos característicos dispersos, usando el algoritmo de SURF (Versión mejorada de SIFT) para encontrar los descriptores de ambas imágenes y posteriormente se utiliza el algoritmo *de K-means* para encontrar la relación entre los puntos característicos de ambas imágenes.

Compararon sus resultados obtenidos con su proceso de alineación planteados contra los resultados obtenidos usando el SDK de Windows para el Kinect v2, mostrando grande mejoría con la métrica *PCK (Percentage of Correct Key-points)*, como se muestra en la Figura 2. 2.

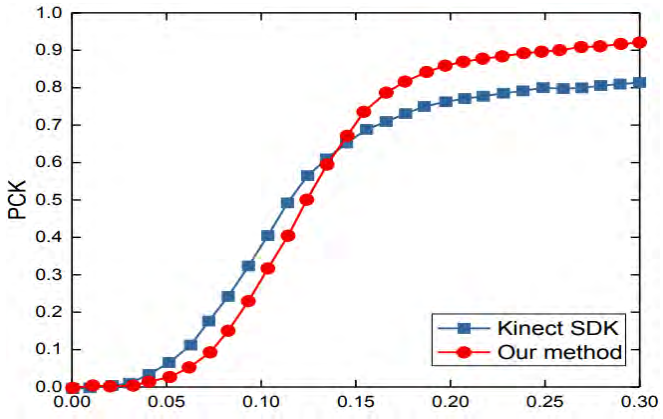

Figura 2. 2 Comparativo de método propuesto para la alineación versus el implementado con el SDK (Tang, Wang and Chen, 2019b)

Como conclusión se resaltan los 3 puntos de innovación.

1.- Realizar la estimación de la pose en 2D, para después complementar la información con la obtenida en 3D.

2.- Utilizar la red *Faster-RCNN* para la detección de humanos, y *Resnet50* para la extracción de características para mejorar la detección de humanos.

3.- Encontrar la correspondencia entre los descriptores de ambas imágenes, imagen de color e imagen de profundidad.

2.1.2 Overlapped Human Pose Estimation using Non-Maximum Suppression based on Shape Similarity (Yu *et al.,* 2022)

En este artículo se explica el uso del enfoque Top-Down, partiendo de la detección del cuerpo de la persona, para luego estimar la pose mediante un mapa de calor.

La principal problemática abordada en este artículo tiene relación con la fase de la detección de la persona, donde las redes neuronales profundas al regresar el cuadro delimitador del cuerpo encontrado, generalmente encarcelan y regresan información no necesaria cuando se encuentran dos o más cuerpos juntos para la estimación de las articulaciones; por lo que proponen minimizar esa información por medio de una supresión de no máximos (NMS por sus siglas en inglés) para filtrar los cuadros delimitadores del cuerpo, se omiten las de menor puntuación.

Para abordar este problema proponen una supresión no máxima basada en la similitud de forma para detectar el cuerpo de la persona en los cuadros delimitadores, para luego extraer las coordenadas de los puntos clave y etiquetar la información para filtrar candidatos con ruido del mapa de calor y lograr una mejor precisión de la estimación.

En el artículo se reportan resoluciones mediante otros métodos que proponen el uso de NMS para abordar el problema eliminando los cuadros delimitadores reduplicados para el mismo objetivo, pero todos los delimitadores alrededor del objetivo mayores al umbral propuesto son suprimidos, incluyendo los que representan prospectos de diferentes objetivos. Por lo que la propuesta de usar NMS basada en la similitud de forma ayuda a distinguir entre los prospectos con sobre cubrimiento, evitando que se eliminen por error al definir un umbral.

En este artículo se menciona que la detección de personas traslapadas entre ellas ya se ha resuelto, pero los cuadros delimitadores de cada persona contienen las articulaciones no solo de la persona principal sino también de las demás personas que comparten un mismo cuadro delimitador.

Las aportaciones principales en este artículo son las siguientes:

1. Implementación de NMS basado en similitud de forma para los cuadros de detección de personas encimados
2. Introducción de un *Transformer* para predecir el mapa de calor de las articulaciones y etiquetarlas.
3. La introducción de un algoritmo de ajuste para corregir las articulaciones encontradas en el cuadro delimitador de la persona, ya que el mapa de calor podría encontrar articulaciones que no pertenecen al cuerpo principal a estimar su pose, causando así una confusión y error en la estimación de la pose.

El flujo de trabajo propuesto en este artículo se muestra en la Figura 2. 3.

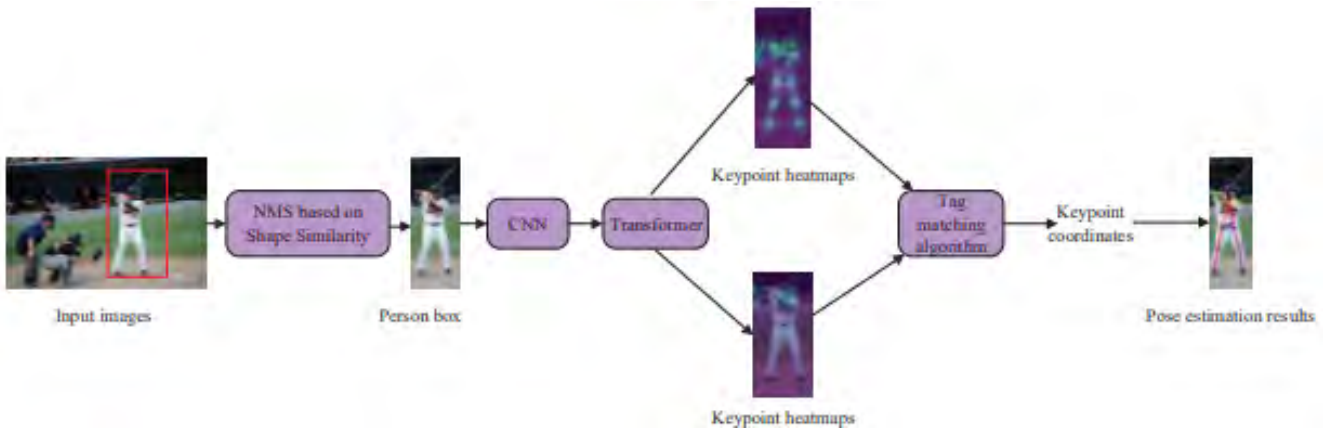

Figura 2. 3 Método propuesto (Top-Down) *(Yu* et al*., 2022)*

2.1.3 Articulated Human Pose Estimation Using Greedy Approach (Kherwa *et al.*, 2021)

En este artículo se cubre un enfoque *Bottom-Up* para la estimación de la pose, esto por medio de una representación no paramétrica, características para localizar las articulaciones para cada individuo.

Entre los aportes y novedades está la introducción de una arquitectura multietapa con dos ramas paralelas, una de las ramas estima las articulaciones del cuerpo a través de puntos de calor, mientras que la otra rama captura la orientación de la arquitectura través de vectores. Posteriormente se asocia las articulaciones encontradas al cuerpo

correspondiente por medio de una técnica que se conoce como *greedy part association vector.* Dónde los vectores 2D pretenden dar la translación codificada de la posición de las articulaciones encontradas y la dirección respectiva de la orientación de las partes del cuerpo.

El flujo de trabajo propuesto es el que se muestra en la Figura 2. 4.

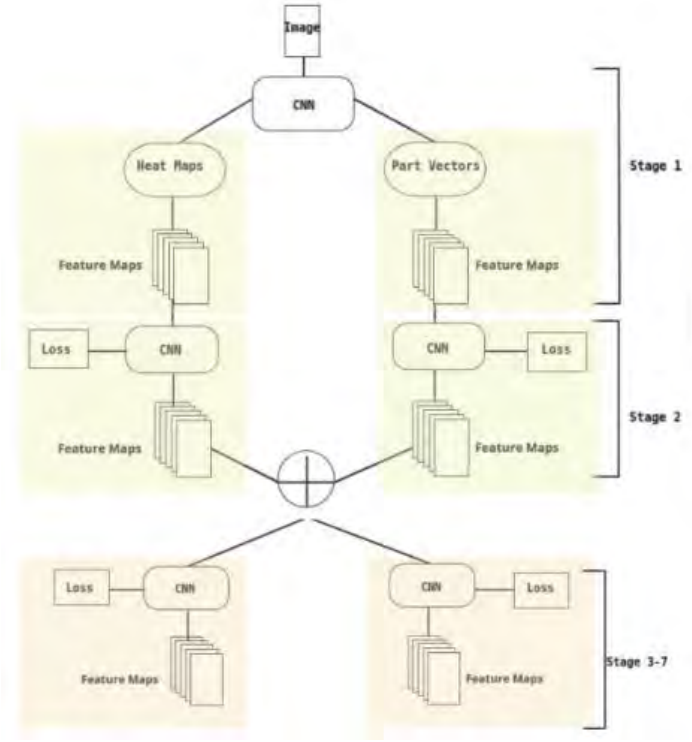

Figura 2. 4 Modelo propuesto *(Kherwa* et al.*, 2021)*

La arquitectura propuesta recibe una imagen de tamaño fijo y produce las anotaciones de la pose en una imagen 2D, esto para cada persona encontrada en la imagen de entrada. En paralelo, la arquitectura atiende dos tareas, la primera consiste en predecir una aproximación para puntos de calor *H* para cada cuerpo en la imagen, y la segunda tarea consiste en predecir un conjunto de vectores 2D que representan la asociación de articulaciones *P* para cada par de diferentes articulaciones.

Se menciona que los puntos de calor producidos por la red neuronal convolucional son altamente confiables, estos puntos de calor son un conjunto de matrices que almacenan la confianza de que cada pixel corresponda a una parte del cuerpo.

El problema planteado en el artículo viene a partir de tener todos los puntos de calor en la imagen, ¿cómo se asocian todas las articulaciones encontradas? Una manera de abordar esto, es por medio de usar una línea geométrica como fórmula de punto medio, pero esto tiende a dar falsas asociaciones cuando la imagen está muy abarrotada de personas. Esto principalmente porque solo se codifica la información de la ubicación de las articulaciones y no la dirección de ellas. La propuesta para solucionar esto es el uso de un enfoque conocido como *greedy part association vector* el cual considera tanto la posición como la orientación a treves del área entera del par de articulaciones.

Los resultados del artículo se evalúan comparando con otros modelos propuestos para el mismo enfoque y con el conjunto de datos de COCO.

2.1.4 TransPose: Towards Explainable Human Pose Estimation by Transformer

(Yang *et al.,* 2020)

En el siguiente artículo se explica el funcionamiento de las redes neuronales convolucionales que estiman la pose, esto surge por la falta de una explicación explicita de cómo se logra la localización de las articulaciones.

La metodología propuesta busca capturar primero la dependencia espacial entre las partes del cuerpo humano y explicar su predicción. La arquitectura propuesta para este modelo se muestra en la Figura 2. 5.

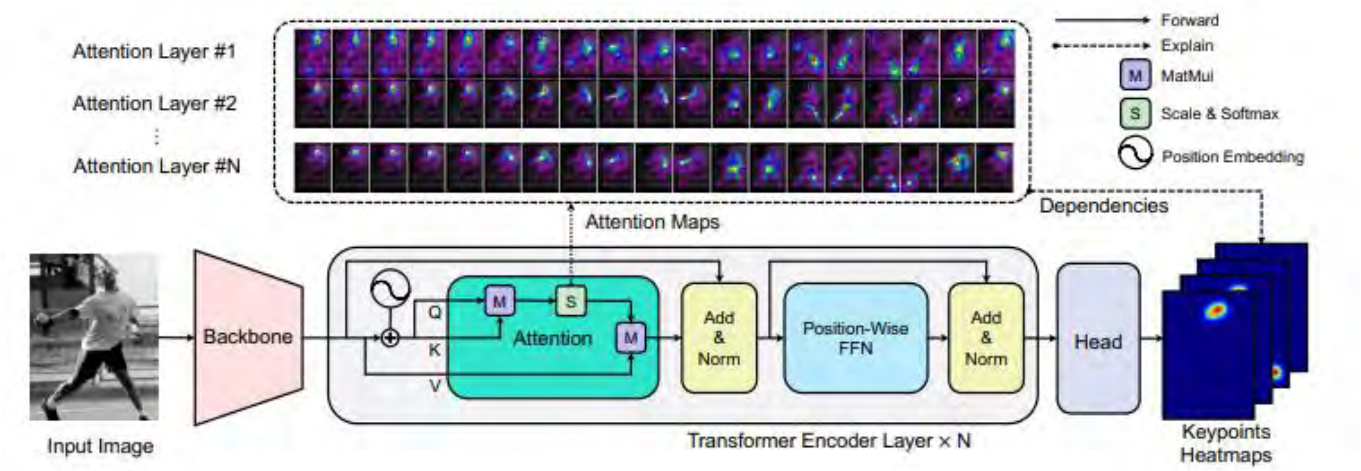

Figura 2. 5 arquitectura propuesta y sus bloques *(Yang* et al*., 2020)*

La arquitectura está compuesta de tres componentes principales: una red neuronal convolucional para la extracción de características de bajo nivel en la imagen; un *Transformer Encoder* para capturar interacciones espaciales de largo rango entre el vector de características a través de sus ubicaciones; una sección denominada *Head* para la predicción del mapa de calor de las articulaciones.

La experimentación se llevó a cabo con el *dataset* de COCO siguiendo un enfoque *Top-Down* con imágenes de una resolución de 256x192. Los resultados obtenidos muestran que un mecanismo de atención es muy prometedor para capturar y explicar las relaciones espaciales dependientes de la imagen.

2.1.5 HPERL: 3D Human Pose Estimation from RGB and LiDAR (Fürst *et al.,* 2020)

En el artículo se menciona que la estimación de la pose se aborda con sensores RGB o RGB-D bajo ambientes controlados, lo que acaba siendo poco efectivo para efectos prácticos en aplicaciones como la estimación de la pose en transeúntes. La detección de transeúntes da lugar a una distancia promedio aproximada de 5 a 50 metros, por lo que

es relevante el rango de distancia al cual trabajan los sensores, a causa de que las cámaras RGB-D carecen de un largo rango de alcance y la estimación de la distancia no es precisa con cámaras RGB, se plantea el uso de sensores LiDAR, los cuales son cada vez más accesibles y más usados para tareas dentro del campo.

En el artículo se plantea lo siguiente para realizar la estimación de la pose 3D más precisa.

- Una arquitectura que funciona con imágenes RGB y la nube de puntos del sensor LiDAR para una mayor precisión.
- Un procedimiento de entrenamiento supervisado débil para una estimación de la pose 2D y 3D simultáneo usando solo etiquetas para las poses 2D.
- Métricas de evaluación para el rendimiento 3D del enfoque sin costosas anotaciones de pose 3D.

Los resultados obtenidos son evaluados en el *dataset* PedX, un *dataset* nuevo y poco usado, por lo que la validación la realizaron con el *dataset* de MPII, usando las métricas más usadas en el estado del arte, PCK, PCKh y la métrica nueva para evaluaciones indirectas en 3D, CDE. La Figura 2. 6 muestra los resultados obtenidos comparados con imágenes RGB y RGB+LiDAR.

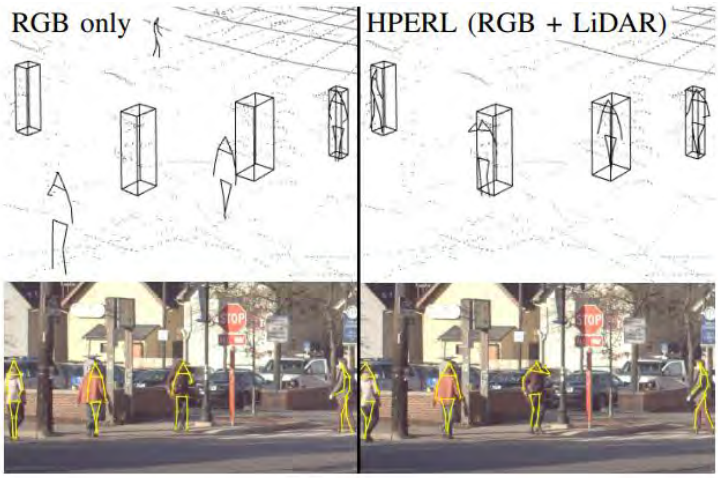

Figura 2. 6 Estimación 3D de la pose con información RGB únicamente y con información de profundidad *(Fürst* et al.*, 2020)*

2.1.6 Bottom-Up Human pose Estimation Via Disentangled Keypoint Regression
(Geng *et al.,* 2021b)

En el siguiente artículo se plantea el uso del enfoque *Bottom-Up* para la estimación de la pose. Se menciona que este enfoque realiza la regresión de los puntos clave para predecir su ubicación perteneciente a cada persona. La regresión de la posición de los puntos clave necesita de manera precisa aprender representaciones que se centren en las regiones de los puntos clave.

Partiendo de ese concepto de regresión se propuso un método llamado DEKR (*DisEntangled Keypoint Regression),* adaptando un proceso de convolución a través de transformaciones espaciales a nivel de pixel para activar los píxeles que se encuentran en las regiones de puntos clave y, a continuación, aprender las representaciones de estos píxeles activados, de modo que las representaciones aprendidas puedan centrarse en las regiones clave.

El enfoque propuesto de regresión directa supera los esquemas de agrupamiento y detección de puntos clave y logra nuevos resultados en el estado del arte con enfoque bottom-up en datasets referentes, COCO y *CrowdPose*.

2.1.7 A Review of Human Pose Estimation from Single Image (Khan and Wan, 2018)

En este artículo se resumen y comentan trabajos asociados a la estimación de la pose a partir de una sola imagen, un análisis que cubra todos los aspectos de este dominio es difícil debido a la diversidad de aplicaciones del área, por lo que solo se enfoca en las contribuciones más importantes en la estimación de la pose humana.

Desde los principales trabajos asociados a la estimación de la pose se puede notar un gran avance hasta los trabajos más recientes. Donde la estimación de la pose con estructuras pictóricas es una de las ideas principales a partir de la conceptualización de cómo los humanos pueden reconocer la pose al mirar la ubicación de las diferentes partes del cuerpo humano.

Esto se propuso como un método compuesto de dos principales módulos; identificación de las partes y las configuraciones de las partes para formar una estructura.

Sin embargo, el uso de Redes Neuronales Profundas (RNP) demostró una mejora relevante y dejo atrás otros enfoques. Partiendo así con un problema de regresión con tres diferentes enfoques basados en las RNP, holísticos basados en partes y una combinación de los dos anteriores.

El uso de estructuras pictoriales en conjunto con redes neuronales muestran un resultado favorable al considerarse observaciones físicas de la estructura de la pose, aun así, la mejoría en la estimación de la pose se ve limitada a problemas como lo son las oclusiones.

2.1.8 A review of deep learning techniques for 2D and 3D human pose estimation (Ben Gamra and Akhloufi, 2021)

El artículo revisa secciones necesarias para abordar la estimación de la pose para los diferentes flujos de trabajos y enfoques propuestos en la literatura como se observa en la Figura 2. 7.

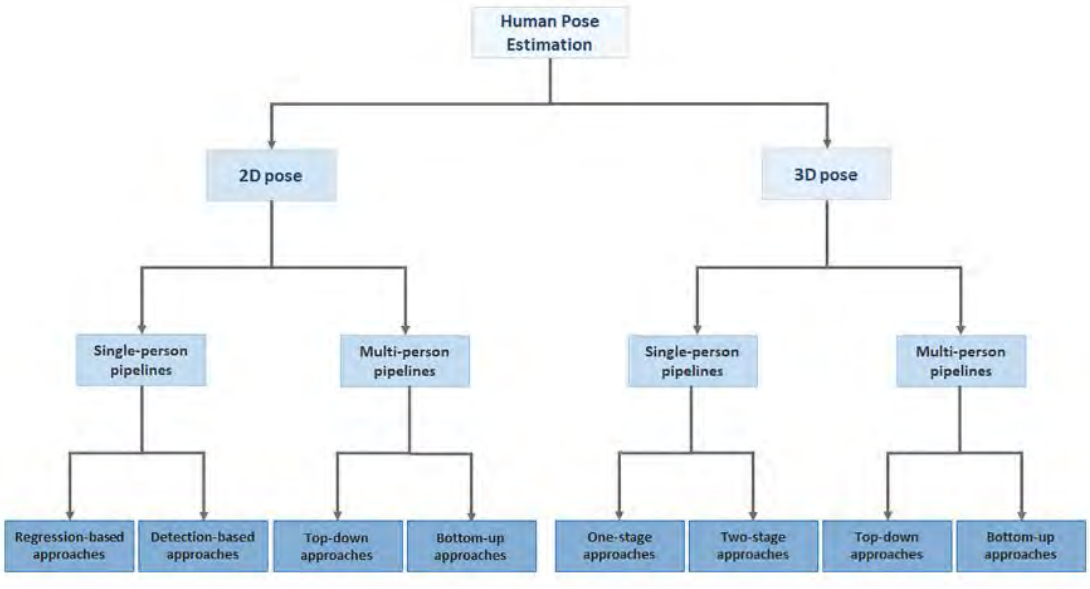

Figura 2. 7 Diferentes enfoques para la estimación de la pose (Ben Gamra and Akhloufi, 2021)

Para estimación de la pose para una única persona en 2D se presentan dos enfoques, uno basado en regresión y otro en la detección.

En la estimación basada en regresión se mencionan varios trabajos que abordan este enfoque; pero igual se menciona que muchos de los trabajos son superados por un enfoque basado en la detección, principalmente porque no se captura el contexto en la regresión.

También se menciona que la estimación de la pose basada en regresión generalmente es abordada con mapas de calor para expresar todas las posibles articulaciones candidatas, posteriormente se realiza un proceso de decodificación que se refiere a transformar los mapas de calor generados a coordenadas de las articulaciones, las cuales son la predicción final.

En la estimación basada en la detección el *ground truth* es generado a partir de la posición de las articulaciones, generalmente aplicando una distribución gaussiana 2D centrada en la ubicación conjunta. Después de la introducción de las redes neuronales convolucionales, la información contextual se captura utilizando capas de convolución.

Los grandes retos se mencionan con este enfoque son: detectar las articulaciones a partir del mapa de calor generado para cada una de ellas y realizar el par de articulaciones correspondientes.

En las conclusiones del trabajo se mencionan los retos de la estimación de la pose, y como cada uno propone diferentes técnicas y enfoques para solucionar pequeños problemas, dejando la oportunidad de trabajar con las combinaciones disponibles de dichos trabajos para generar nuevos enfoques o mejores resultados.

2.1.9 2D Human Pose Estimation based on Object Detection using RGB-D information (Park, Ji and Chun, 2018)

El artículo se propone un nuevo enfoque para la estimación de la pose en 2D basado en información RGB-D, con un enfoque *Top-Down* partiendo de un proceso de segmentación para el fondo.

La información de profundidad es combinada con la información RGB existente para resolver el problema de oclusión y reconocimiento debido a la falta de información topológica de la imagen 2D. La información de profundidad se puede obtener mediante el cálculo de la disparidad entre las imágenes generadas desde las cámaras izquierda y derecha y generando el mapa de profundidad.

Se realiza la detección de objetos segmentando un objeto utilizando información de profundidad en una región segmentada por información RGB.

Posteriormente, la región del objeto detectado es re-escalada para generar los datos de entrada para la estimación de la pose. Esta información o datos generados son aplicado al modelo propuesto, CPN (*Convolutional Pose Machine)* el cual genera una predicción secuencial basada en redes neuronales convolucionales para las articulaciones. El proceso descrito hasta ahora se muestra en la Figura 2. 8.

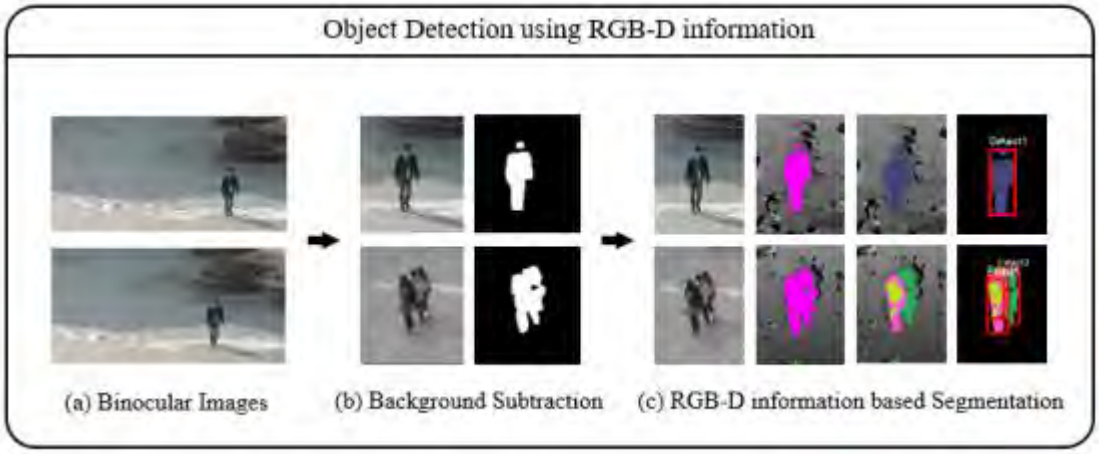

Figura 2. 8 Flujo de trabajo propuesto para la estimación de la pose (Park, Ji and Chun, 2018)

El proceso de estimación de la pose se divide entonces en dos secciones, donde a partir de la imagen segmentada se procede a calcular un mapa de confianza para cada articulación, localizando en total 14 articulaciones; posteriormente se procede a generar el modelo de la pose identificada. Estos pasos se muestran en la Figura 2. 9.

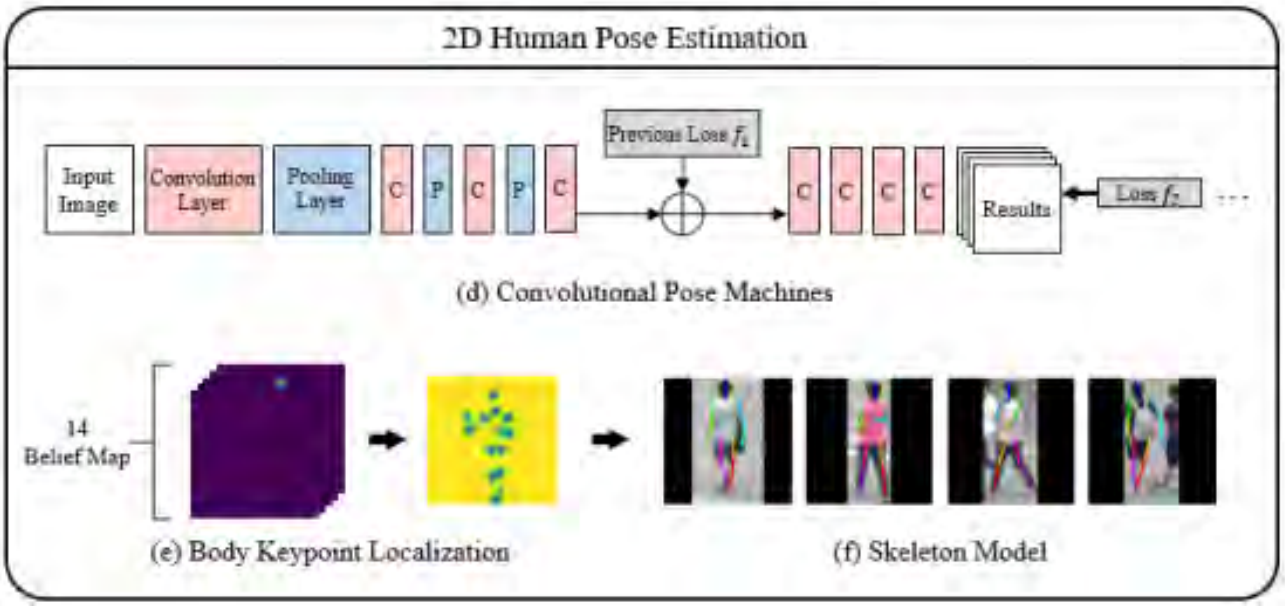

Figura 2. 9 Pasos para la estimación de la pose con el modelo planteado (Park, Ji and Chun, 2018)

En el artículo se logró adquirir información de la profundidad para generar regiones de interés. Por consiguiente, la pose humana se puede estimar produciendo un modelo de cuerpo humano representado por un esqueleto modelo utilizando 14 puntos clave de posiciones conjuntas humanas.

2.1.10 Low-resolution human pose estimation (Wang *et al.*, 2022)

El artículo plantea que la estimación de la pose huma es abordada a una alta resolución en su mayoría de ocasiones. La alta resolución es asociada en el artículo a 256x256 pixeles, mientras que una baja resolución es considerada a 128x96 pixeles. También se especifica que imágenes de alta resolución brindan más información para la estimación

de la pose, mientras que la captura del cuerpo humano se realiza a una baja resolución por la variabilidad de escalas a la que se puede llegar a presentar la persona en la imagen.

El problema de usar imágenes de alta resolución es que conlleva un mayor coste computacional en memoria y tiempo de ejecución.

En el artículo se plantea que la estimación de la pose basada en mapas de calor es más sensible a errores en la detección de articulaciones, debido a la poca cantidad de pixeles para cada articulación.

Se propone, según el estado del arte relacionado con la investigación, usar un aprendizaje ajustado, que ha mostrado ser un enfoque efectivo para estimar la pose.

En la Figura 2. 10 se muestra el flujo de trabajo de un estimador de la pose basado en ajuste.

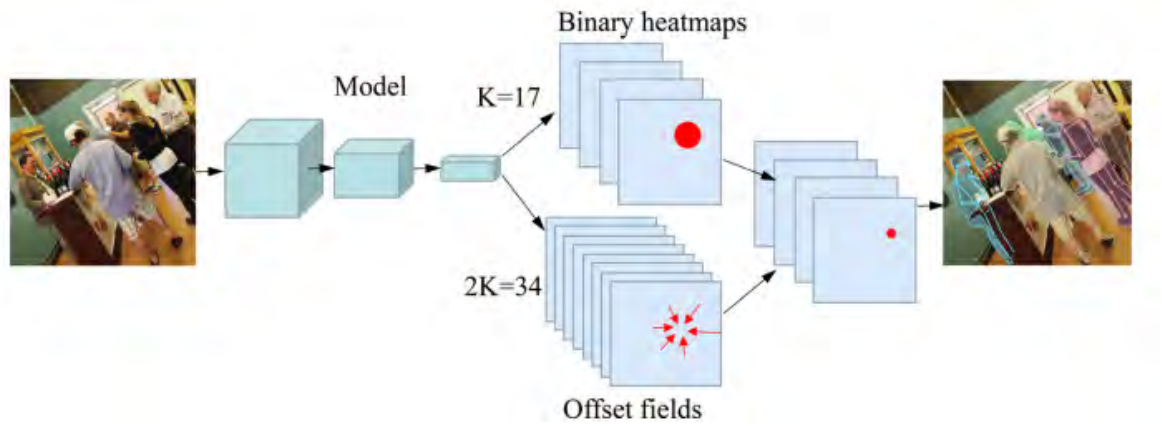

Figura 2. 10 Arquitectura propuesta *(Wang* et al*., 2022)*

Los experimentos se hicieron con datos de COCO, demostrando superioridad sobre otros trabajos asociados con baja resolución.

2.2 Modelos de estimación de la pose

De los modelos de estimación de la pose que existen, se han considerado y estudiado algunos de los referidos dentro del estado del arte, así como los mencionados a continuación.

2.2.1 Stacked Hourglass Networks for Human Pose Estimation (Newell, Yang and Deng, 2016)

En el artículo se propone una nueva arquitectura de red neuronal con un mecanismo de supervisión para las capas medias de la red. El origen de esta propuesta viene dada por la necesidad de mejorar el rendimiento de las redes neuronales.

El diseño de la arquitectura es referido como "*Stacked Hourglass"* y está basada en pasos sucesivos de *pooling* y *upsampling* que son realizados para producir un conjunto de predicciones de las articulaciones encontradas en la imagen.

El diseño puede capturar información en cada escala, ya que la evidencia local es esencial para identificar características como rostros en cada escala. Una estimación final de la pose requiere de un entendimiento del cuerpo completo*,* como la orientación de la persona, el agrupamiento entre articulaciones y así como la relación entre articulaciones adyacentes. Estas características son mejor reconocidas en diferentes escalas de la imagen. El diseño *Hourglass* es un diseño simple que tiene la capacidad de capturar todas esas características a diferentes escalas y reunirlas para generar la predicción de las articulaciones.

La configuración de la arquitectura *Hourglass* es la siguiente:

1. Capas de convolución y *max pooling* son usadas para procesar las características más bajas a una resolución baja.
2. Una vez alcanzada la resolución más baja, la red comienza la secuencia *top-down* de muestreo ascendente y la combinación de características a diferentes escalas.

3. Cuando se alcanza la resolución de salida, dos rondas consecutivas de convolución de 1x1 son aplicadas para producir la predicción final de la red.

La predicción de la red es un mapa de calor para cada articulación que predice la probabilidad de la presencia de una articulación para cada pixel.

El número de *hourglasses* en la red tiene un desempeño directo en los resultados, mientras mayor sea el número de ellos en la red, mejor será el resultado. Así como la disminución del número de estas resultará en un crecimiento en la capacidad de cada uno individualmente.

En la Figura 2. 11 se muestra el resultado del cambio en la predicción de la pose desde el segundo *hourglass* hasta el octavo implementado en la red.

Figura 2. 11 Resultados obtenidos a diferentes niveles de la capa (Newell, Yang and Deng, 2016)

La implementación de esta arquitectura demostró su eficiencia para la estimación de la pose por medio de la supervisión intermedia

2.2.2 YOLO-POSE: Enhancing YOLO for Multi Person Pose Estimation Using Object Key Point Similarity Loss (Debapriya M, Soyeb N and Manu M, 2022a)

En este artículo se plantea el uso del detector de objetos YOLO para la estimación de la pose. La problemática sobre la cual se parte es que algoritmos que abordan la detección de la pose a partir del enfoque del uso de mapas de calor para las articulaciones tienden a ser clasificados como subóptimos debido a que no son entrenados de extremo a extremo y el entrenamiento depende en la sustitución de la función de pérdida L1, lo cual no es el equivalente a maximizar la métrica de evaluación, por ejemplo, la métrica de OKS (*Object Keypoint Similarity*).

El artículo menciona cómo la mayoría de los modelos abordan un enfoque *top-down* o *bottom-up*, mencionando sus ventajas y desventajas de ambos, sin embargo, dentro de la propuesta está un modelo que utiliza ambos enfoques, así como también aborda el problema de estimación de la pose sin el uso de mapas de calor para las articulaciones y varios procesos de post procesamiento no estandarizados.

Las atribuciones de manera precisa son las siguiente:

- Solucionar la estimación de la pose para múltiples personas de manera lineal con la detección de objetos, abordando problemas de múltiples escalas en las personas y problemas de oclusión.
- Un enfoque libre de mapas de calor con un post procesamiento enfocado a la detección de objetos en vez de usar un enfoque más complejo normalmente utilizado con los mapas de calor, como lo es NMS, ajustes de escala, regresión, etc.
- Extender la idea del uso de la función IoU (Intersección sobre unión, en español) para la detección de las articulaciones con cuadros delimitadores.

En el artículo se menciona como el uso del enfoque *top-down* tiene problemas cuando se parte de una localización errónea de la persona, en la Figura 2. 12 se puede ver cómo los resultados obtenidos por YOLO-Pose no se ven afectados por esta mala localización.

Figura 2. 12 Resultado de estimación de la pose con enfoque Top-Down mejorado por YOLO-Pose (Debapriya M, Soyeb N and Manu M, 2022a)

Que una articulación no esté en el cuadro delimitador causa que en la etapa de detección de las articulaciones no haya manera de encontrar las articulaciones que estén fuera del cuadro delimitador. El enfoque *Bottom-up* puede lidiar con esta desventaja, sin embargo, tiene una mayor complejidad su implementación.

El modelo fue evaluado con el conjunto de datos de COCO, y demostró ser competitivo con la métrica mAP comparado con otros trabajos del estado del arte con enfoque bottom-up. De igual manera demostró un fuerte resultado para la métrica AP50, pasando todos los trabajos en este mismo enfoque *bottom-up*.

2.3 Estimación de la pose con oclusiones

En esta sección se menciona alguno de los retos ya previamente mencionados en la literatura sobre las oclusiones causadas por entornos complejos durante la estimación de la pose humana, así como algunas de las técnicas propuestas en el estado del arte para abordar la problemática.

2.3.1 Quantification of Occlusion Handling Capability of a 3D Human Pose Estimation Framework (Ghafoor and Mahmood, 2022)

El artículo presenta un enfoque a los problemas de oclusión ya estudiados, donde se sabe que la estimación de la pose humana en 3D es mayormente abordada a partir de la estimación de la pose en 2D. Métodos estudiados en este artículo demuestran que no siempre manejan el problema de oclusiones de manera directa, mientras que algunos otros métodos han propuesto algunos mecanismos de detección de la pose conscientes de la oclusión; sin embargo, la capacidad de estos métodos no se ha cuantificado a fondo.

Se propone un método que aborde de manera explícita la falta de articulaciones y determine la capacidad del manejo de las oclusiones. Este método se comparó con otro método ya existente en el estado del arte en términos de la métrica *Mean Per Joint Position Error* (MPJPE).

Las contribuciones principales en el artículo son las siguientes:

1. Un marco guiado por oclusión es propuesto basado en una red neuronal convolucional de dilatación temporal para la estimación de la pose en la presencia de oclusiones severas
2. La capacidad del manejo de las oclusiones por las redes neuronales profundas es cuantificada a partir de la pérdida de articulaciones en la imagen de entrada 2D que ya tiene las articulaciones identificadas. Se quitan algunas articulaciones o todas las uniones en los primeros cuadros de la secuencia. La calidad de las poses estimadas es cuantificada usando la métrica MPJPE, así como la precisión del reconocimiento de la acción.
3. Análisis comprensivo de las evaluaciones en los *datasets* públicos como, Human 3.6M, NTU RGB-D y SYSU, donde se demuestra mejoramiento significable para la estimación de la pose en 3D y en la precisión del reconocimiento de la actividad.

El flujo de trabajo propuesto es el que se muestra en la Figura 2. 13.

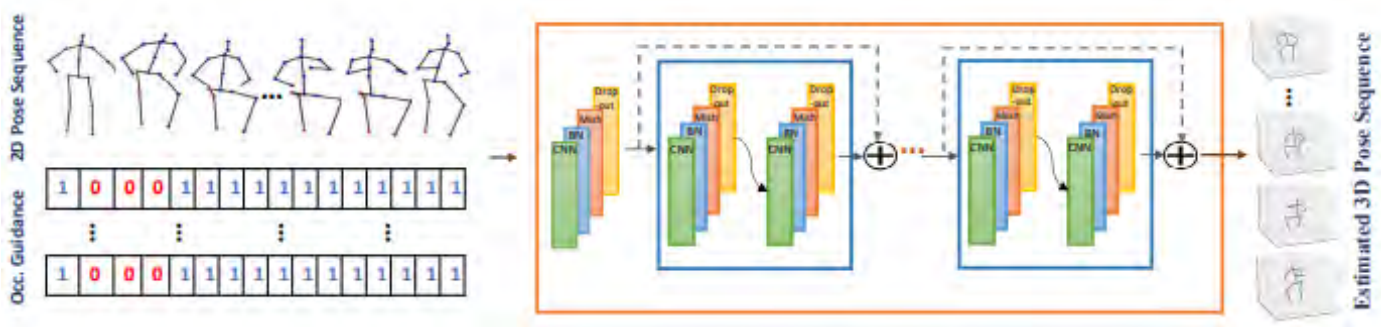

Figura 2. 13 Mecanismo propuesto para manejar la oclusión (Ghafoor and Mahmood, 2022)

Donde el mecanismo de oclusiones etiquetadas es parte del aporte principal. A partir de la predicción de la pose en 2D, donde la mayoría de las articulaciones son estimadas en conjunto a un mecanismo de confianza, este mismo mecanismo de confianza puede ser utilizado para crear la matriz de guía de oclusiones donde los valores de 0 corresponden a articulaciones con una baja confianza y 1 a una alta confianza, de esta manera se provee información de las articulaciones faltantes a la red.

Se reporta que le método mapea exitosamente la relación de la información 2D como entrada en conjunto con la matriz de guía de articulaciones ocluidas a un espacio de representación 3D de salida.

2.3.2 Identifying Visible Parts via Pose Estimation for Occluded Person Re-Identification (Miao, Wu and Yang, 2021)

En el artículo se aborda el problema de la oclusión en personas re-identificadas (re-id), que es uno de los principales retos en escenarios de recuperación de personas del mundo real. Métodos anteriores en el problema de re-identificación ocluida generalmente asume que solo las extremidades están ocluidas, eliminando así las oclusiones mediante el recorte manual. Sin embargo, esto no siempre se mantiene en la práctica.

Para el desarrollo de la experimentación todas las imágenes del conjunto de consultas están ocluidas, mientras que el conjunto de galería contiene imágenes ocluidas y no ocluidas, lo que es más desafiante y práctico. Se propone beneficiar a la *re-id* ocluida por puntos de referencia de pose en tres aspectos.

Primero, se utiliza la información de ubicación espacial de los puntos de referencia visibles para filtrar el ruido de las regiones de oclusión.

En segundo lugar, se utilizan puntos de referencia visibles para generar la incrustación de la pose, que se utiliza como puertas de canal para recalibrar las características del canal.

En tercer lugar, en las pruebas, las características de pieza comúnmente visibles se utilizan para la comparación. Además, se construyó un conjunto de datos de *re-id* ocluido a gran escala, *Occluded-DukeMTMC.*

El flujo del método propuesto se muestra en la Figura 2. 14.

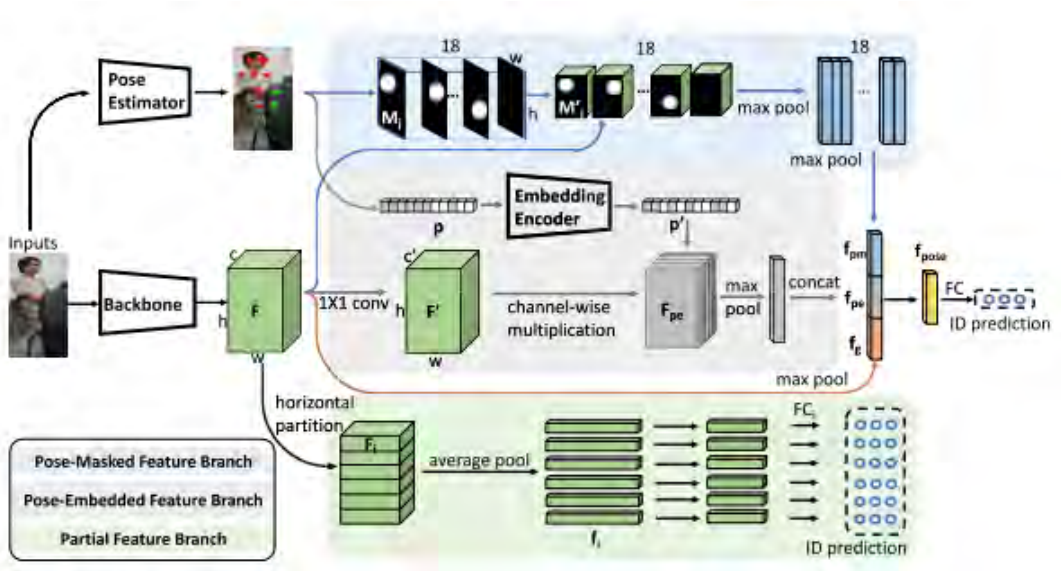

Figura 2. 14 Arquitectura propuesta (Miao, Wu and Yang, 2021)

2.3.3 Realistic Augmentation for Effective 2d Human Pose Estimation Under Occlusion (Ansarian and Amer, 2021a)

En este trabajo se discute la efectividad de la estimación de la pose humana considerando el reto dado por las oclusiones que ocurren con alta frecuencia en imágenes en aplicaciones reales, basándose en el uso de modelos *CNN* para que se ponga mayor atención a patrones globales y no tanto a locales.

Se plantea el desarrollo de un conjunto de datos con aumento de oclusiones de manera artificial, a través de la inserción de oclusiones a partir del cuadro delimitador de la persona y anexando objetos reales dentro del cuadro como se muestra en la Figura 2. 15.

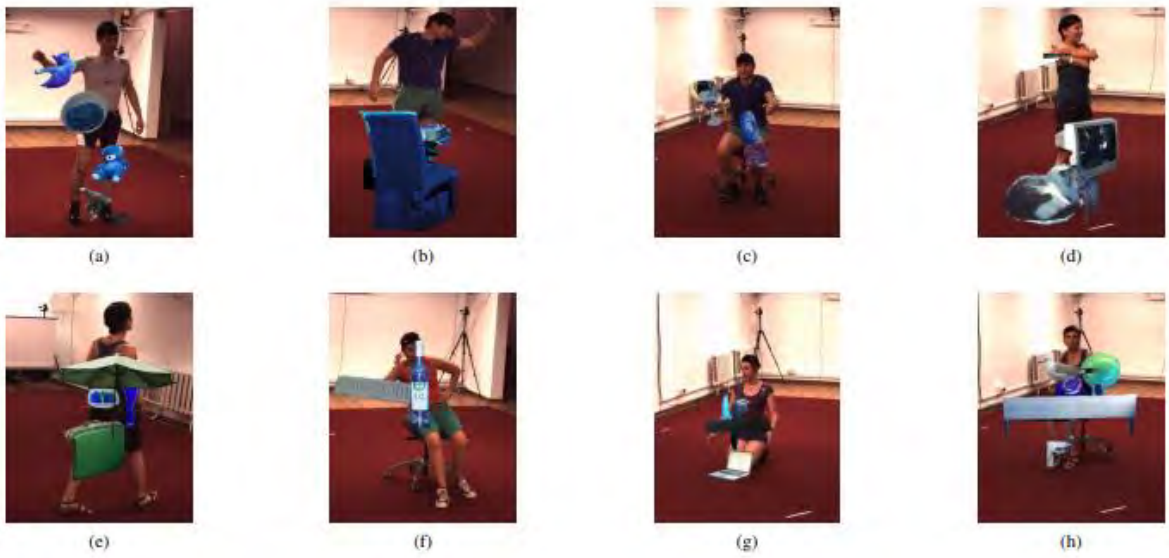

Figura 2. 15 Oclusiones sintéticas realistas (Ansarian and Amer, 2021b)

La evaluación del conjunto de datos se implementó a partir de la validación de la estimación de la pose con la métrica MPJPE en el conjunto de datos Human3.6m sin oclusiones y posteriormente con oclusiones, dando como resultado una mejoría notable sobre la inferencia del modelo basado en CNN entrenado con oclusiones sintéticas, sobre esta problemática.

2.4 Reconocimiento de actividades humanas

Dentro del estudio del reconocimiento de actividades humanas se ha encontrado un amplio uso de la estimación de la pose humana como una nueva metodología basada en

la extracción de puntos característicos del cuerpo humano, sin embargo, estos no son los únicos enfoques reportados en la literatura.

En el campo del reconocimiento de actividades humanas se han presentado trabajos de investigación mediante diferentes métodos y problemáticas específicas para el reconocimiento de las actividades. En (Ghosh et al., 2020) se trabaja con el uso de dispositivos inteligentes para el reconocimiento de actividades a través del procesamiento de datos inerciales con algoritmos de aprendizaje automático. En un estudio (Shinde et al., 2018) se hace uso de cámaras con la finalidad de identificar la actividad estática por medio del procesamiento completo de la imagen con el uso de redes neuronales convolucionales y aprendizaje profundo con YOLO, mientras que Kellokumpu et al. (2005) realiza la sustracción del fondo para segmentar la imagen y usar la segmentación de la persona como un descriptor invariante para usarlo con máquinas de vector soporte y determinar la pose de la persona. Yadav et al. (2021) trabaja en conjunto con ambos dispositivos, en un sistema denominado multimodal usando tanto el procesamiento en cámaras como con los dispositivos inteligentes, de la misma manera en (Ehatisham-Ul-Haq et al., 2019) se detalla un sistema multimodal que utiliza información de diversos sensores, como lo son cámaras RGB, sensores de profundidad y sensores inerciales vestibles. Así mismo reporta un mejor resultado en la integración de estos para la clasificación de las actividades. Nadeem et al. (2020) trabaja con el reconocimiento de las actividades basado en la detección de partes del cuerpo a partir de la segmentación del fondo y localización de puntos característicos básicos. Kilbas et al. (2022) trabaja únicamente con un conjunto de puntos característicos para calcular las distancias entre estos, y así entrenar un modelo basado en redes neuronales para la clasificación de 36 poses.

2.5 Antecedentes

En el CENIDET se encuentran varios trabajos de investigación ya realizados y activos con similitud al trabajo realizado en esta tesis, por lo que en esta sección se reporta una breve síntesis de los trabajos realizados previamente.

- “Caracterización automática de una muestra de acciones en vídeo” (María Luisa Dávila García, 2009). En este trabajo se desarrolló un sistema de visión artificial capaz de caracterizar una muestra automática de acciones humanas en vídeo. La caracterización numérica consiste en medidas geométricas entre regiones de piel en movimiento y zonas de referencia durante la secuencia. Las principales técnicas utilizadas fueron substracción de fondo para detectar movimiento y segmentación de imágenes por color para extraer regiones de piel. Se realizó una interpretación de la caracterización numérica utilizando autómatas. Con la herramienta GrafO3D fue posible visualizar el resultado de la caracterización gráficamente y comprobar el buen funcionamiento del sistema.
- “Caracterización Visual de Movimientos “Sospechosos” de Personas en Estacionamientos” (Rafael Alcantar Juárez, 2013). En este trabajo se desarrolló un SVI (Sistema de Videovigilancia Inteligente), capaz de hacer la detección y seguimiento de personas, y en segundo lugar hacer la caracterización de cinco movimientos sospechosos: correr, levantar una o ambas manos, merodear, agacharse y acostarse. El sistema tiene tres módulos, el primero es el procesamiento digital de imágenes, las tareas de generar y actualizar el modelo de fondo, la segmentación de objetos en movimiento, la supresión de sombras, la validación de objetos como personas con filtro de tamaño, la localización de partes del cuerpo y la tarea de seguimiento. En el segundo módulo se calculan propiedades geométricas y descriptores útiles para caracterizar los movimientos propuestos; finalmente, el tercero gestiona la caracterización, determina y alerta de un evento sospechoso. Las pruebas se realizaron con dos tipos de conjuntos de videos; con bancos de videos disponibles en internet citados en el estado del arte, cuya principal característica es que son grabados en escenarios poco complejos. Y con videos de un estacionamiento adquiridos personalmente, en un ambiente natural sin control.
- “Caracterización de Imágenes en Movimiento: Correr y Caminar” (Jorge Alfredo Saldaña García, 2007). En este trabajo se desarrolló un sistema de visión artificial, que realiza de manera automática, la extracción de características de los

movimientos corporales al caminar y correr, considerando partes del cuerpo no estudiadas como lo es el ángulo de la espalda y la posición de codos y manos. Respetando siempre el movimiento natural del sujeto, mediante un procedimiento no invasivo, en un ambiente controlado. Para la caracterización se aplicaron los algoritmos See5 y Cubist, cuyo modelo resultante arroja nueva información sobre las partes del cuerpo que intervienen en los movimientos bajo estudio. Los modelos se validaron usando algoritmos de reconocimiento de patrones de libre acceso en internet y por lo tanto reconocidos y probados.

2.6. Discusión

La estimación de la pose, que consiste en determinar la posición y orientación de un conjunto de puntos característicos en un espacio tridimensional. Desempeña un papel fundamental en una gran variedad de aplicaciones, desde la manipulación de robots hasta la realidad aumentada. En los últimos años, se ha realizado una cantidad significativa de investigaciones, explorando diversas técnicas y algoritmos como los que se mencionan en la sección 2.1, esto para mejorar la robustez de la estimación de la pose.

Al revisar los trabajos previos, se observa una amplia variedad de enfoques utilizados para abordar la estimación de la pose, desde enfoques en 2D y 3D hasta la estimación *Botton-Up* y *Top-Down* como lo menciona (Ben Gamra and Akhloufi, 2021). sin embargo, las metodologías van más allá de esta clasificación, llegando a utilizar una metodología de estructuras pictoriales para restringir la relación entre puntos característicos como lo menciona (Belagiannis *et al.,* 2016). Así mismo (Yu *et al.,* 2022) implementa técnicas como la supresión de no máximos para descartar puntos característicos con baja puntuación en escenarios de traslape u oclusiones, siendo así el uso de técnicas de aprendizaje profundo cada vez más utilizadas para la estimación de la pose 2D y 3D respectivamente (Dang *et al.,* 2019a; Huang, Huang and Tang, 2021).

Esto ha puesto en resalte las problemáticas presentes con la estimación de la pose, como lo menciona (Ghafoor and Mahmood, 2022), la falta de métricas para cuantificar la eficiencia de la estimación de la pose manejando oclusiones es evidente en los modelos actuales, sin embargo, existen varias métricas para validar da fiabilidad de los modelos de

estimación de la pose, como las mencionadas en (Munea *et al.,* 2020), donde se mencionan métricas como *PCP, PDJ, PCK y PCKh.*

De igual manera la estimación de la pose humana con oclusiones ha sido abordada con la implementación de oclusiones artificiales, esto por la falta de conjuntos de datos que caractericen este problema del mundo real como lo menciona (Ansarian and Amer, 2021a), y proponiendo un conjunto de datos basado en el aumento de datos con oclusiones, acercando así los estudios a datos más reales.

CAPÍTULO 3 ANÁLISIS TEÓRICO

3.1 Metodología abordada

En esta sección se presenta la metodología abordada para el desarrollo del proyecto con la finalidad de dar cumplimiento a los objetivos previamente planteados. La metodología está dividida en tres etapas principales, la primera de ellas es la etapa de localización, asociada a la ubicación de la persona o personas en la imagen y posteriormente a la ubicación de los puntos característicos de cada persona en la imagen.

Se realizó una búsqueda del estado del arte de las técnicas disponibles, donde se desarrolló un documento en forma de resumen donde se encuentran alrededor de 10 técnicas y modelos disponibles para abordar esta primera etapa, así una vez identificado dichos modelos se realizó otro análisis experimental con ayuda de la herramienta *"MM Pose"* para determinar qué modelo de estimación de la pose incorporar al sistema en la etapa de localización.

Para la segunda etapa correspondiente a la clasificación de la pose se planteó un conjunto de procesos para entrenar un modelo de aprendizaje automático, este modelo fue seleccionado a partir de la experimentación realizada para comparar el rendimiento de al menos cuatro modelos, los cuales fueron *lineal regression, random forest, ridge classifier y gradient boosting classifier.*

Una vez entrenado el modelo para la clasificación de la pose se realizó la tercera etapa, la cual corresponde a la inferencia del modelo en conjunto con las dos primeras etapas, para lo cual se creó una interfaz de usuario que permite mayor flexibilidad entre el sistema

y el usuario. Dicha interfaz abarca desde la selección, entrenamiento e inferencia de los modelos.

La metodología explicada de manera breve previamente se muestra en la Figura 3. 1.

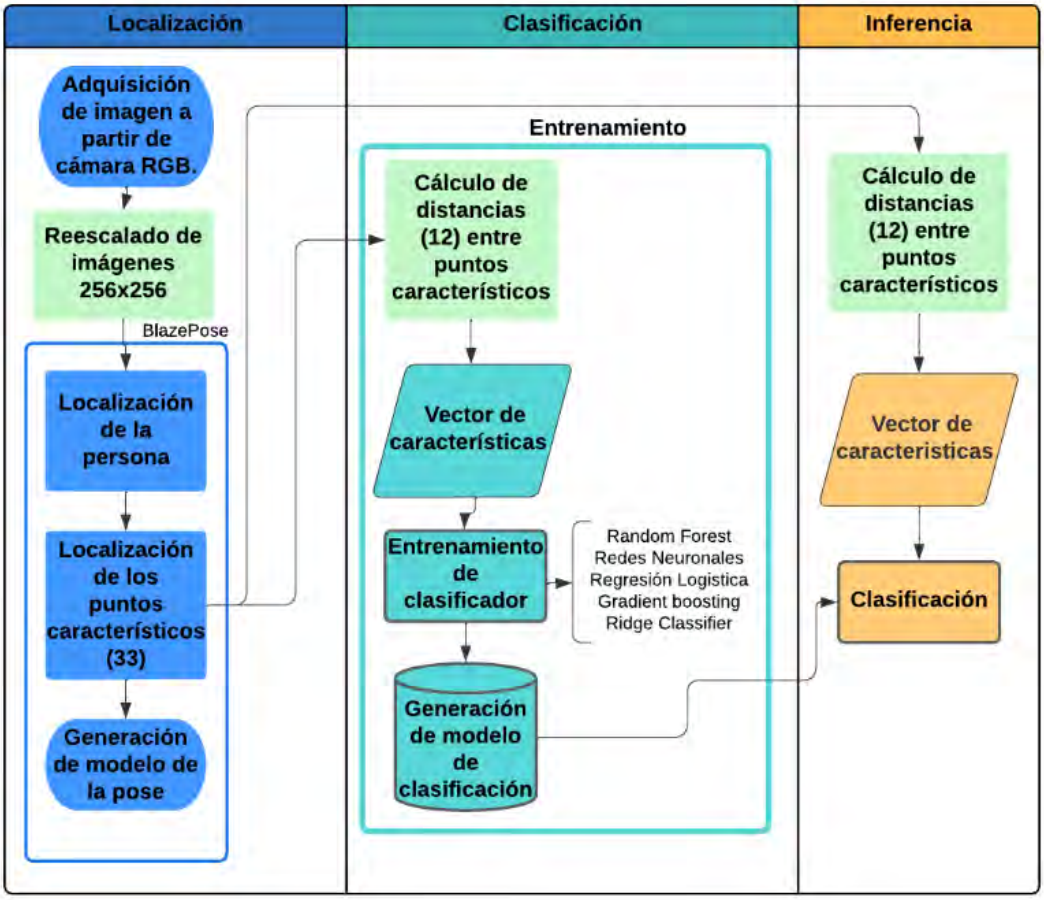

Figura 3. 1 Diagrama de metodología abordada

A continuación, se describe de manera detallada cada sección del diagrama presentado en la Figura anterior.

Localización

La etapa de localización está conformada por tres procesos principales, la localización de las personas, la localización de los puntos característicos y la generación del modelo de la pose.

- La adquisición de la imagen es realizada a partir del uso de una cámara RGB, por lo cual se limita a el modelo de localización a trabajar con imágenes en el espacio de colores RGB y con una resolución de 480x640 pixeles.
- La localización de las personas se abordó de manera automática por medio de los modelos de estimación de la pose, los cuales ya tienen integrada la etapa de localización de la persona como paso indispensable del modelo, sin embargo, es importante mencionar que esto puede diferir según el enfoque que se utilice con los diferentes modelos. Más detalle de esto en la sección 3.1.1.
- La localización de los puntos característicos igualmente es abordada de manera directa por el modelo de estimación de la pose, sin embargo, es importante mencionar ciertos criterios para la consideración de la localización de un punto característico como válido. En esta etapa el modelo toma como entrada la sección delimitada por el localizador de personas para buscar los puntos característicos y regresar un vector que contiene la ubicación cartesiana de cada uno de ellos, así como información adicional que representa la probabilidad o visibilidad de cada uno de ellos.
- La generación del modelo de la pose realiza la proyección de dicha pose utilizando un modelo cinemático previamente definido.

Clasificación

Para la etapa de clasificación se propusieron 4 modelos de aprendizaje automático, para la evaluación de cada uno de ellos con los diferentes conjuntos de datos, con la finalidad de identificar cuál mostró mejor rendimiento para el problema específico de clasificación de la pose.

- La etapa del cálculo de distancias tomó el vector de las ubicaciones de los puntos característicos y un subconjunto de estos puntos para realizar el cálculo de doce distancias. Esto ayuda a tener mayor robustez al modelo de clasificación de la pose con respecto a la variabilidad de las poses a rotaciones y translaciones.
- Las distancias son calculadas a partir de la distancia manhattan y posteriormente guardadas como vector de características para el entrenamiento de los modelos de clasificación.
- La etapa de entrenamiento es realizada a través Python con la librería de *scikit-learn*.
- Una vez entrenado cada modelo y evaluado, se seleccionó el mejor modelo para ser utilizado como modelo final en el sistema.

Inferencia

En la etapa de inferencia se creó todo el sistema con base en los modelos de estimación de la pose y de clasificación, para lo cual se desarrolló en Python una interfaz gráfica que ayuda a visualizar el resultado de la inferencia de ambos modelos, más detalles de la interfaz se plantean en la sección **3.5. Interfaz gráfica**.

Como resultado final de la metodología implementada se desarrolló un diagrama de flujo con la solución conceptual de la Figura 3. 2.

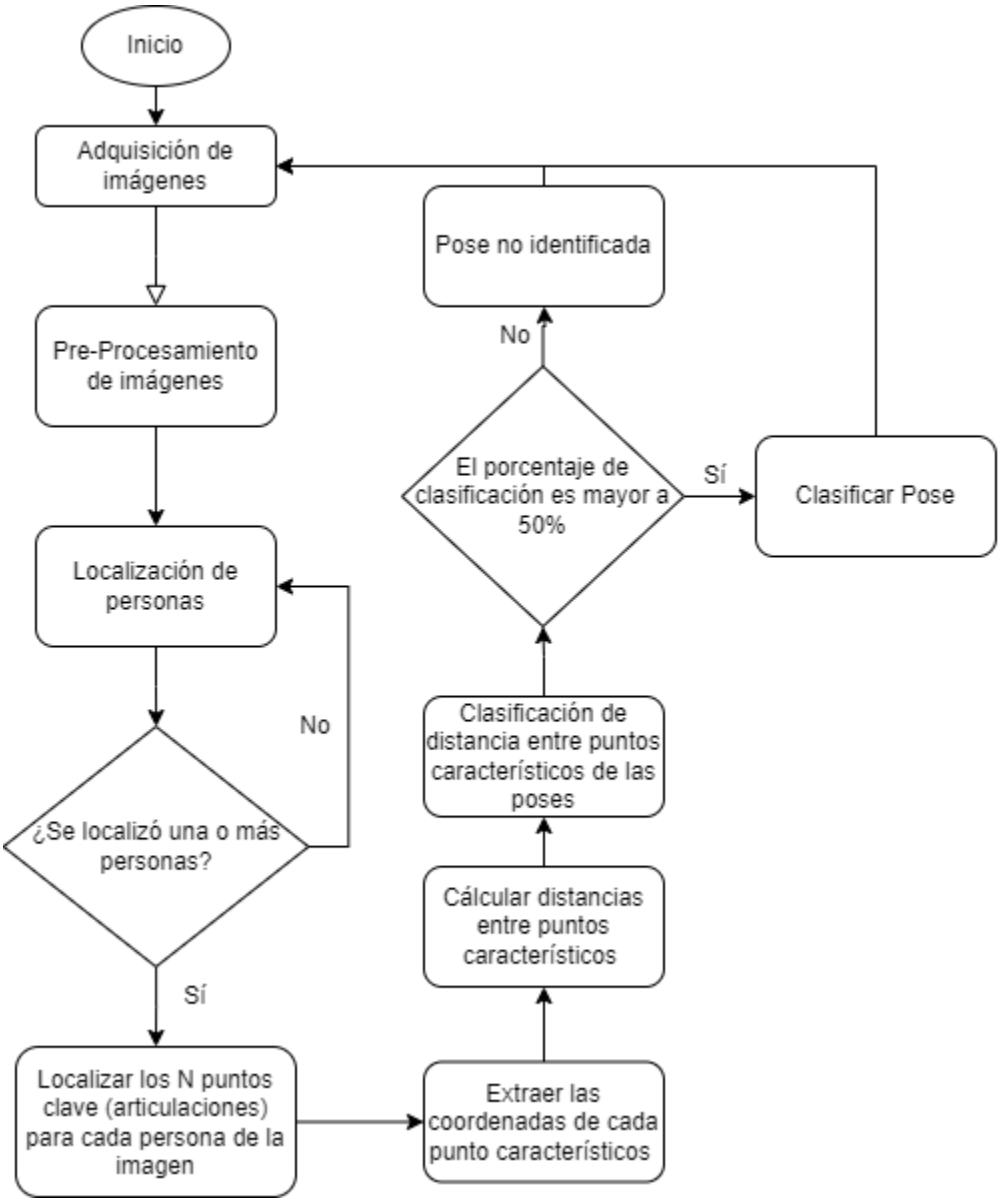

Figura 3. 2 Diagrama de flujo de solución conceptual

3.2. Recursos utilizados

Para el desarrollo de la experimentación se utilizaron los siguientes recursos:

- 1 laptop Dell G15 con procesador Intel Core i5, tarjeta gráfica (GPU) Geforce gtx 1080, 16GB de memoria RAM
- 1 Cámara Logitech C920 HD Pro Webcam, 1080P FULL HD 1080p/30 fps

3.3. Bibliotecas utilizadas

Para el desarrollo del modelo se han utilizado diferentes bibliotecas disponibles en Python, en esta sección se mencionan las librerías utilizadas en la Tabla 3. 1, así como su versión.

Tabla 3. 1 Bibliotecas usadas

Librería	Versión
Python	3.10
Numpy	1.21.6
opencv-python	4.6
pandas	1.4.3
scikit-learn	1.1.2
seaborn	0.12.1
torch	1.12.1+cu116
torchvision	0.13.1+cu116
Custom TKinter	8.6.12

3.4. Conjunto de datos

Para la validación del modelo de estimación de la pose humana se han usado dos subconjuntos de imágenes extraídas de OcHuman y CrowdPose.

CrowdPose

Este conjunto de datos que contiene 20,000 imágenes y un total de 80,000 anotaciones de poses humanas con 14 puntos característicos, el conjunto de datos está creado con la finalidad de proveer imágenes con presencia de abarrotamientos humanos, donde se presentan mayores problemas de oclusión y variabilidad de escalas de las personas (Li *et al.*, 2019a).

En la Figura 3. 3 se muestran algunos ejemplos de imágenes extraídas de *CrowdPose*

Figura 3. 3 Imágenes en CrowdPose *(Li* et al*., 2019b)*

OcHuman

Este conjunto de datos se centra en oclusiones fuertes presentes en humanos con anotaciones comprensivas que incluyen no solo las anotaciones de los puntos característicos, sino también de los cuadros delimitadores de personas y su segmentación (Zhang et al., 2019).

Algunas de las imágenes del conjunto de datos de OcHuman se muestran en la Figura 3. 4.

Figura 3. 4 Imágenes de OcHuman *(Zhang* et al*., 2019b)*

Conjunto de datos de la literatura

En la literatura se identificó el conjunto de datos propuesto por (Kilbas, Gribanov and Paringer, 2022a), el cual está constituido por 35 clases asociadas a varios ejercicios y poses comunes.

El conjunto de datos fue recolectado de manera interna capturando un gran conjunto de poses realizadas por 4 mujeres y 3 hombres. Las personas proveen una cantidad moderada de variabilidad con respecto a su complexión corporal.

La lista completa de todas las poses presentadas en el conjunto de datos se muestra en la Tabla 3. 2.

Tabla 3. 2 Lista de clases

Id Clase	*Nombre de clase*	*Número de ejemplos*
1	Manos en el pecho	428
2	Manos detrás de la espalda, frente	433
3	Manos detrás de la espalda, trasera	301
4	Tijeras, frontal	582
5	Tijeras sobre la cabeza, frontal	613
6	Tijeras sobre la cabeza, trasera	593
7	Sentadilla, frontal	239
8	Sentadilla, lateral izquierda	304
9	Sentadilla lateral derecha	266
10	Tocando hombro izquierdo con mano derecha	260
11	Tocando hombro derecho con mano derecha	186

Tabla 3. 2 Lista de clases (continuación)

Id Clase	*Nombre de clase*	*Número de ejemplos*
12	Tocando hombro izquierdo con mano izquierda	147
13	Tocando hombro derecho con mano izquierda	115
14	Manos detrás de cabeza, frontal	334
15	Manos detrás de cabeza, trasera	252
16	Manos encimadas, mano derecha encima, frontal	426
17	Manos encimadas, mano derecha encima, trasera	362
18	Manos encimadas, mano izquierda encima, frontal	454
19	Manos encimadas, mano izquierda encima, trasera	259
20	Tocando tobillo derecho con mano derecha	93
21	Tocando tobillo izquierdo con mano derecha	60
22	Tocando tobillo izquierdo con mano izquierda	86
23	Tocando tobillo derecho con mano izquierda	49
24	Tocando rodilla derecha con mano derecha	175
25	Tocando rodilla izquierda con mano derecha	91
26	Tocando rodilla izquierda con mano izquierda	146
27	Tocando Rodilla derecha con mano izquierda	58
28	Doblando pierna izquierda hacia la pelvis, frontal	190
29	Doblando pierna izquierda hacia pelvis, trasera	197
30	Doblando pierna derecha hacia pelvis, frontal	151
31	Doblando pierna derecha hacia pelvis, trasera	199
32	Doblando brazo derecho al codo por atrás, frontal	426
33	Doblando brazo derecho al codo por atrás, trasera	320
34	Doblando brazo izquierdo al codo por atrás, frontal	368
35	Doblando brazo izquierdo a altura de codo	238
36	Poses intermedias	12295

El conjunto de dato creado a partir de las imágenes consiste de la extracción de los puntos característicos del cuerpo por medio de un modelo de estimación de la pose, el cual es implementado de manera interna, sin dar detalle a profundidad, únicamente muestra la distribución y representación del modelo a través de la Figura 3. 5.

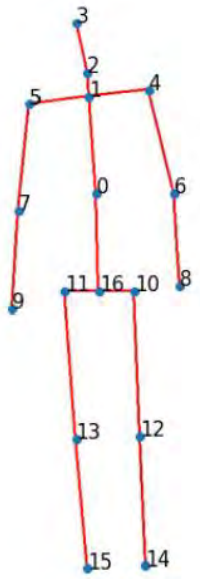

Figura 3. 5 Ejemplo de pose (modelo de estimación de la pose)

Conjunto de datos propuesto

El conjunto de datos propuesto consiste en 10 clases seleccionadas bajo los siguientes criterios:

- Poses atípicas. – Poses con menor número de ocurrencia en los conjuntos de datos de entrenamiento de los modelos de estimación de la pose como se menciona en el estudio realizado por (Hwang, Yang and Kwak, 2020).
- Poses que presentan mayor complejidad en (Kilbas, Gribanov and Paringer, 2022a) debido a que incurren con pérdida de información por sobreposición entre puntos característicos (oclusión).
- Poses con mayor número de puntos clave ocluidos, como se plantea en el estudio realizado por (Cao *et al.*, 2017b).

Las diez clases seleccionadas bajo los criterios previamente mencionados se encuentran en la **¡Error! No se encuentra el origen de la referencia.**

Tabla 3. 3 Clases de conjunto de datos propio

Id de clase	*Clase*	*Número de ejemplos*
1	Tocando tobillos	526
2	Tocando rodillas	440
3	Tocando rodillas manos cruzadas	375
4	Tocando tobillos manos cruzadas	419
5	Hincado	708
6	Sentado	1052
7	Sentado en piso	510
8	Sentado en piso piernas cruzadas	850
9	Parado	943
	Total de poses	5823

Para la construcción del conjunto de datos se utilizó el modelo de estimación de la pose disponible a través de *MediaPipe*, el cual consiste en 33 puntos característicos descritos por 4 variables, *X, Y, Z* y *Visibilidad*. En la Figura 3. 6. se muestra el modelo de estimación de la pose utilizado.

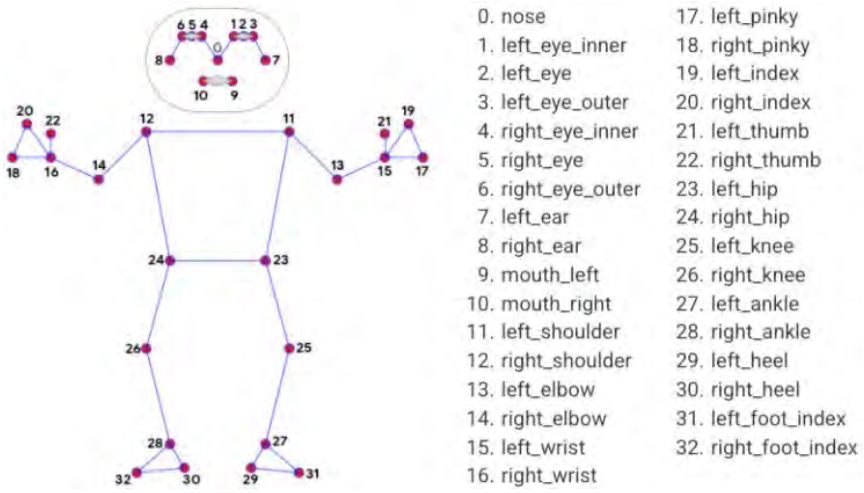

Figura 3. 6 Modelo de estimación de la pose a través de MediaPipe (Bazarevsky and Grishchenko, 2020a)

3.5. Interfaz gráfica

Para la integración final de todos los elementos del proyecto se ha creado una interfaz gráfica de usuario con ayuda de la biblioteca Tkinter en Python. En la Figura 3. 7 se muestra la interfaz gráfica, la cual se encuentra dividida en tres secciones principales, la parte de entrenamiento, inferencia y validación.

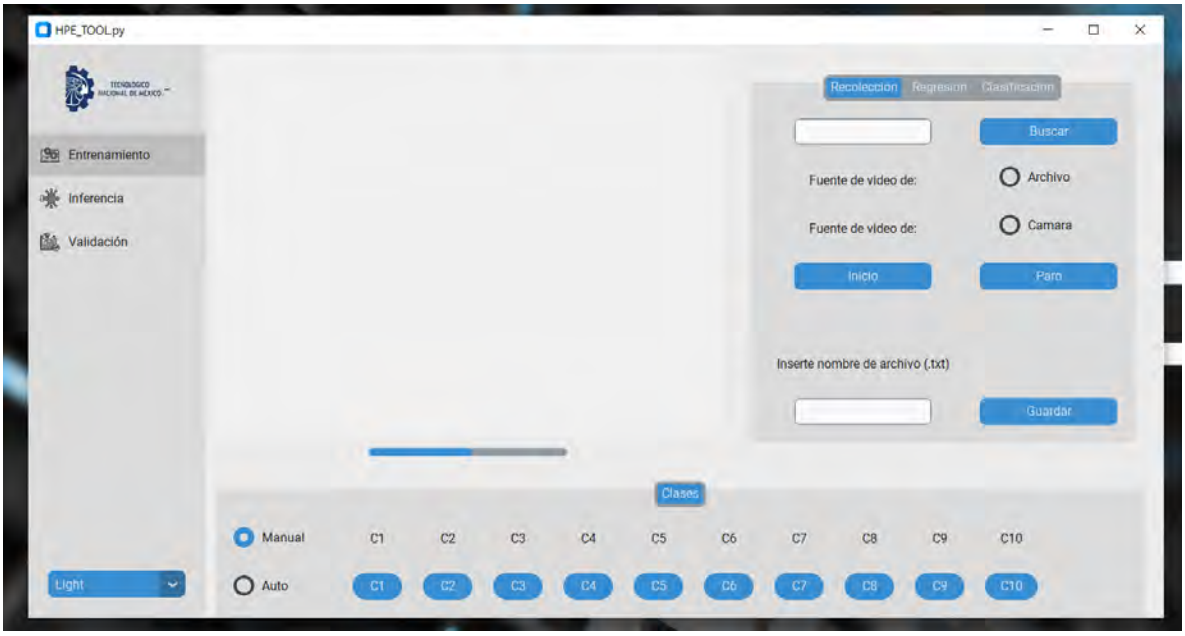

Figura 3. 7 Captura de la interfaz gráfica en la sección de entrenamiento

El diseño de la interfaz consiste de múltiples elementos que ayudan a realizar tanto el proceso de entrenamiento como validación e inferencia, por lo tanto, agiliza el desarrollo de aplicaciones basadas en la clasificación de la pose.

En la Figura 3. 8 se muestra el panel de la sección de entrenamiento, el cual está constituido por múltiples subsecciones para llevar a cabo todas las etapas requeridas para el entrenamiento del modelo de clasificación, desde la selección del modelo de estimación de la pose (4 opciones de modelos). Posteriormente, realizar la recolección de los puntos característicos de manera manual o automática para generar como máximo un conjunto de datos con 10 clases.

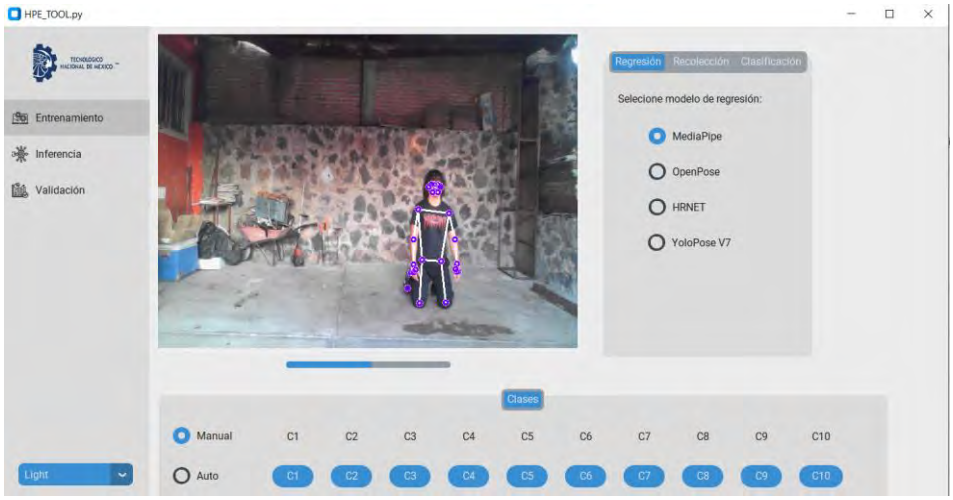

Figura 3. 8 ventana de clasificación para selección de modelo de regresión.

La Figura 3. 9 muestra las ventanas de recolección y clasificación respectivamente, donde en recolección se puede seleccionar el origen del video de entrada, tanto desde un archivo local (MP4) o a través de la cámara del dispositivo local. De igual manera permite seleccionar la ubicación del archivo con las coordenadas generado a partir de la recolección, este archivo generado es un csv o txt posteriormente utilizado para el entrenamiento del modelo de clasificación.

Figura 3. 9 Ventana de clasificación para selección de entrada de video y recolección de datos, así como para la selección de modelo de clasificación

En la Figura 3. 10 se aprecia la ventana de inferencia donde se permite seleccionar el origen del video, ya sea de la cámara del dispositivo o a través de un video cargado del fichero de la computadora, también permite seleccionar el modelo de estimación de la pose, como lo es MediaPipe, OpenPose, HRNet y YoloPose, cabe mencionar que el modelo seleccionado de estimación de la pose debe coincidir con el modelo seleccionado para la recolección de los datos.

Figura 3. 10 Ventana de regresión, selección de origen de video, selección de modelo de estimación de la pose y modelo de clasificación

CAPÍTULO 4 EXPERIMENTACIÓN Y RESULTADOS

4. Experimentación y resultados

La experimentación y resultados presentes en esta sección se encuentran divididos en 3 experimentos realizados.

Experimento 1: El primer experimento consistió en la evaluación de los modelos de estimación de la pose disponibles en la literatura en dos de los conjuntos de datos más asociados al objetivo del problema (Estimar la pose humana con presencia de oclusiones), estos conjuntos de datos son *Ochuman y ClowdPose.*

Experimento 2: El objetivo de este segundo experimento consistió en la evaluación de cuatro modelos de clasificación, esto con la finalidad de identificar qué modelo cumplía con un mejor resultado para la clasificación del conjunto de poses propuesto únicamente con presencia de auto oclusiones.

Experimento 3: El tercer caso experimental consistió en implementación de oclusiones parciales a los videos del conjunto de datos propuesto, esto con la finalidad de evaluar su rendimiento y determinar el modelo de clasificación a incorporar al sistema final.

4.1. Experimento 1: validación de modelos de estimación de la pose con conjuntos de datos OcHuman y ClowdPose

Para determinar qué modelo de estimación de la pose incorporar al sistema final se realizó una validación del funcionamiento de los modelos disponibles, *MediaPipe, HRNet, YoloPose V7 y OpenPose* sobre un subconjunto de 200 imágenes extraídas de *OcHuman y ClowdPose*.

La experimentación se realizó en Python con ayuda de Open MMLlab y Pytorch para cargar y ejecutar los modelos. Se requirió realizar ajustes a cada modelo de estimación de la pose para poder evaluarlos como se plantea y corresponde en cada conjunto de datos.

Para la evaluación con *ClowdPose*, las anotaciones para la evaluación consisten en 14 puntos característicos como se muestran en la Figura 4. 1 y una resolución de 480px x 640px en cada imagen, por lo que fue necesario modificar la distribución de los puntos característicos para cada modelo, así como la resolución de la imagen.

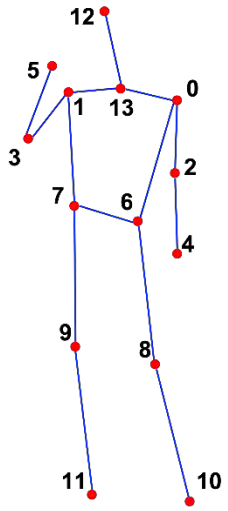

Figura 4. 1 Anotaciones para CrowdPose

Los archivos de configuración necesarios disponibles en el sitio web oficial de Open mmlab, para el caso del conjunto de datos de Crowdpose son los mostrados en la sección de ANEXOS, así como la estructura general utilizada para la carga del conjunto de datos.

Los resultados de la evaluación de los modelos se muestran en la Tabla 4. 1, donde se aprecia que Yolo v7 tiene un buen resultado sobre el promedio de la precisión por encima de los otros modelos, mientras que en caso contrario OpenPose tiene el resultado más bajo con un promedio de precisión de 0.493%.

Tabla 4. 1 Resultados de modelos de estimación de la pose en subconjunto de imágenes de CrowdPose

	AP	*AP50*	*AP75*
HRNet	0.475	0.585	0.529
OpenPose	0.493	0.543	0.372
YoloV7	**0.692**	**0.813**	**0.691**
BlazePose	0.664	0.805	0.713

Para la evaluación de OcHuman las anotaciones igual corresponden a 14 puntos característicos y las imágenes son de diferente resolución, así que se tuvo que adaptar de igual manera los modelos a lo mostrado en la Figura 4. 2.

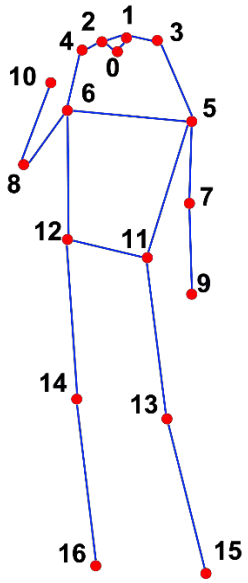

Figura 4. 2 Anotaciones para OcHuman

Los resultados obtenidos con el conjunto de datos de OcHuman se muestran en la Tabla 4. 2, en este conjunto de datos en particular se presenta una disminución en todos los modelos con respecto a su promedio de precisión, esto puede ser debido a que el conjunto de datos representa mayor reto con respecto a las oclusiones. De igual manera Yolo V7 tiene un buen rendimiento con respecto a los resultados con un promedio de precisión de 0.531%.

Tabla 4. 2 Resultados de modelos de estimación de la pose en subconjunto de imágenes de OcHuman

	AP	*AP50*	*AP75*
HRNet	0.425	0.525	0.509
OpenPose	0.422	0.489	0.453
YoloV7	**0.531**	**0.623**	**0.571**
BlazePose	0.483	0.539	0.512

Cabe mencionar que los resultados obtenidos para cada conjunto de datos únicamente se obtuvieron sobre el subconjunto aleatorio de 200 imágenes, por lo que no representa un valor absoluto para los conjuntos de datos. Sin embargo, para el fin de la validación de los modelos se logró mostrar las principales ventajas de cada uno de los modelos de estimación, notándose que para ambos conjuntos de datos el modelo con mayor AP fue YOLOV7, aunque los resultados igual se pueden ver afectado por las modificaciones realizadas de los modelos con respecto a su distribución de puntos característicos y las métricas utilizadas; esto principalmente para modelos que no comparten la misma distribución de puntos característico en el modelo basado en esqueleto. De igual manera la utilización de la métrica de similitud OKS para el cálculo de la precisión promedio permite mayor holgura para valores de 50% y 75% dentro de la proximidad aceptable con relación a los puntos clave reales.

Con base en los resultados y lo mencionado en la literatura donde (Mroz *et al.,* 2021) hace una comparativa de calidad entre BlazePose y OpenPose en un entorno clínico, OpenPose es considerablemente mejor en la estimación de la pose por cuadro, sin embargo, debido a que OpenPose, así como YoloPose y HRNet realizan la estimación de la pose en cada cuadro de video estos tienden a tener fluctuaciones con respecto a la estimación previa, dejando así con mejor opción a BlazePose para la estimación de la pose en videos (Cao *et al.,* 2017a; Bazarevsky and Grishchenko, 2020b; Debapriya M, Soyeb N and Manu M, 2022b)

Igualmente considerando que *BlazePose* se mantuvo como segundo lugar con los resultados obtenidos en los dos conjuntos de datos, por ende, se decidió realizar la incorporación de este modelo de estimación de la pose al sistema final.

4.2. Experimento 2: validación de modelos de clasificación de la pose en conjuntos de datos propios y de la literatura (solo auto oclusiones)

El objetivo de esta experimentación fue demostrar la eficiencia de los modelos de aprendizaje automático para la clasificación de la pose siguiendo la metodología

implementada, de igual manera, seleccionar un modelo de clasificación para la incorporación al sistema final.

Para dicha experimentación se consideró la implementación de cuatro modelos de aprendizaje automático: *random forest, linear regression, gradient boosting classifier y ridge classifier.* Para su implementación y evaluación se utilizó Python y la librería de *sckit learn*, así como librerías auxiliares para el manejo de datos y preprocesamiento de imágenes.

Para la evaluación de los modelos se utilizaron los dos conjuntos de datos con una distribución de 80% para entrenamiento y 20% para validación cada uno, únicamente considerando auto oclusiones, se muestran algunos ejemplos en la Figura 4. 3.

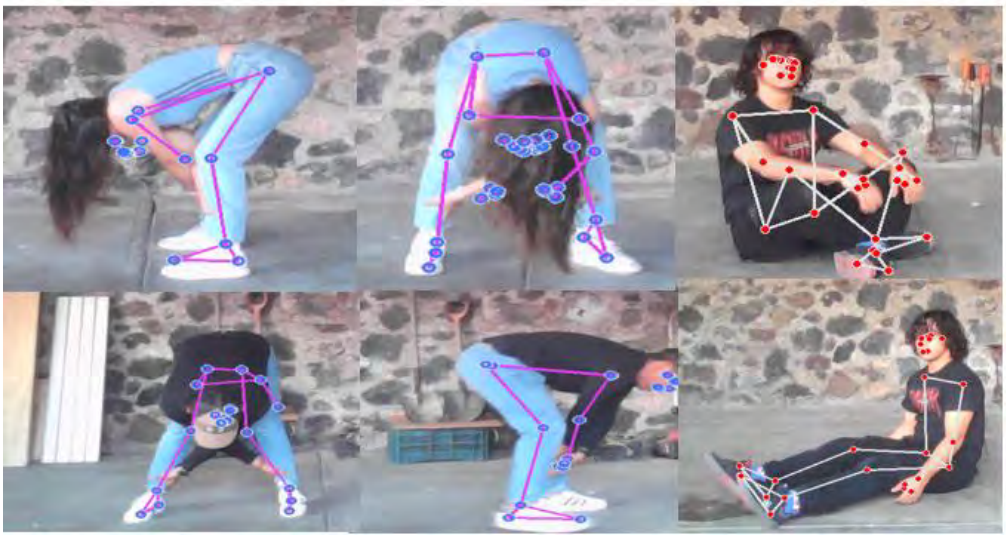

Figura 4. 3 Imágenes de muestra del conjunto de datos propio

Cabe reiterar que no se cuentan con las imágenes del conjunto de datos de (Kilbas, Gribanov and Paringer, 2022b).

Conjunto de datos propios

La implementación de los modelos de clasificación con el conjunto de datos propios arroja un resultado bastante similar para cada uno de los modelos como se muestra en la Tabla

4. 3, donde el modelo con mejores resultados fue *Random Forest* con una exactitud de 91%. Es importante mencionar que, del conjunto de datos utilizado, previamente definido en la tabla 3.4

Tabla 4. 3 Resultados de modelos de clasificación de la pose en conjunto de datos propio

	Logistic Regression	***Ridge Classifier***	***Random forest Classifier***	***Gradient Boosting Classifier***
Exactitud	0.8697	0.8858	**0.9171**	0.8957
Precisión	0.8958	0.8889	**0.9172**	0.8930
Sensibilidad	0.8957	0.8887	**0.9171**	0.8729

El conjunto de datos utilizado es el previamente mencionado en la Tabla 3.4, con un total de 5,823 ejemplos repartidos entre las 9 clases. Los modelos fueron entrenados en un subconjunto representativo del 70% de los datos para el entrenamiento de los modelos de clasificación, dejando así 30% de los datos para la validación.

Conjunto de datos de literatura

Los resultados obtenidos para la clasificación de la pose utilizando el conjunto de datos de la literatura se muestran en la Tabla 4. 4, donde se aprecia que el mejor modelo de clasificación fue *random forest* con una exactitud del 88%, mientras que el clasificador de *ridge classifier* tuvo el peor desempeño. La cantidad de muestras utilizadas para el entrenamiento de los datos consta de 21,696 datos como se describe en la tabla 3.2.

Tabla 4. 4 Resultados de clasificación de la pose en conjunto de datos de (Kilbas, Gribanov and Paringer, 2022a) sin uso de distancias

	Logistic Regression	***Ridge Classifier***	***Random forest Classifier***	***Gradient Boosting Classifier***
Exactitud	0.8475	0.6805	**0.8819**	0.8245
Precisión	0.8476	0.5978	**0.8820**	0.8230
Sensibilidad	0.8475	0.6805	**0.8819**	0.8221

Con la implementación del cálculo de distancia entre los puntos característicos para los modelos se puede observar un aumento en la precisión del modelo como se muestra en la Tabla 4.4 para el caso del conjunto de datos propio y en la Tabla 4. 5 para el de la literatura.

El cálculo de las distancias se realizó como se muestra en la Figura 4. 4, donde se obtuvo un vector de características con únicamente 12 elementos, correspondientes a las distancias entre los puntos seleccionados.

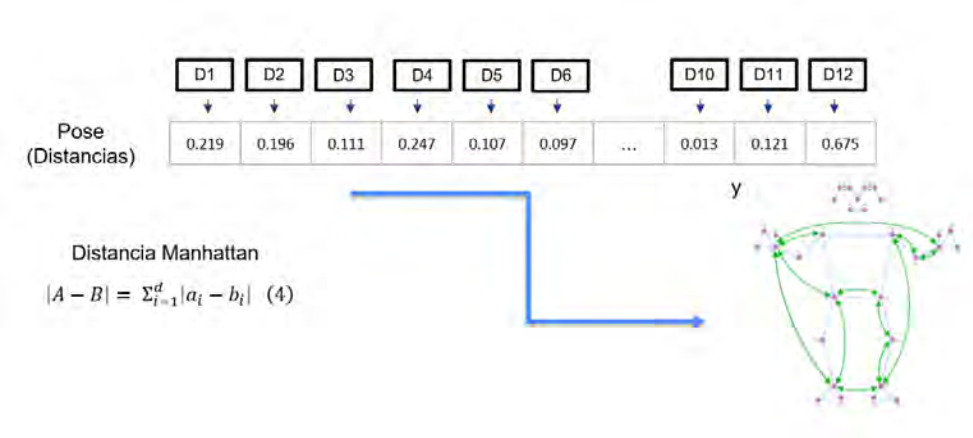

Figura 4. 4 cálculo de distancias

Para el caso de la Tabla 4.5, se aprecia que el algoritmo que mejor rendimiento tuvo fue el del clasificador *gradient boosting con un 98.9%, siendo* ahora *random forest* el algoritmo con un menor desempeño.

Tabla 4. 5 resultados de los modelos de clasificación de la pose en conjunto de datos propios a través del cálculo de distancias

	Logistic Regression	***Ridge Classifier***	***Random forest Classifier***	***Gradient Boosting Classifier***
Exactitud	0.9845	0.9450	0.9313	**0.9891**
Precisión	0.9849	0.9451	0.9313	**0.9892**
Sensibilidad	0.9845	0.9450	0.9313	**0.9891**

Tabla 4. 6 Resultados de los modelos de clasificación de la pose en conjunto de datos de la literatura a través del cálculo de distancias

	Logistic Regression	***Ridge Classifier***	***Random forest Classifier***	***Gradient Boosting Classifier***
Exactitud	0.8675	0.5807	**0.8911**	0.8345
Precisión	0.8476	0.5975	**0.8820**	0.8120
|Sensibilidad	0.8671	0.5805	**0.8913**	0.8221

Los modelos de la Tabla 4.6 muestran un buen resultado para la clasificación de las poses estáticas, por lo que se puede confiar en su desempeño para esta tarea. Pero para la descripción de las poses es meticuloso porque existe una gran variabilidad intraclase para cada pose, y esto también se ve afectado por la variabilidad en la captura de las poses desde diferentes perspectivas, pero la codificación de la información a través de distancias ayuda a la robustez de los modelos a la variabilidad intraclase y rotaciones.

4.3. Experimento 3: validación de modelos de clasificación de la pose en conjuntos de datos propios y de la literatura (con oclusiones parciales)

El objetivo de este caso de experimentación es exponer el modelo de estimación de la pose como el de clasificación a oclusiones sintéticas parciales, durante el proceso de entrenamiento y de inferencia del modelo. Por ende, se han generado oclusiones parciales sobre los conjuntos de datos de entrenamiento en las imágenes para el caso del conjunto propio.

Cabe mencionar que los modelos de estimación de la pose cuentan con mecanismos para identificar la probabilidad de que un punto característico realmente sea la coordenada a la que se está señalando, o igualmente se puede considerar como la visibilidad del punto característico como lo planteqa *BlazePose*, por ende se puede identificar a través de esta variable la confianza o exactitud de la estimación y ayuda a identificar si algún punto característico asociado a alguna articulación está ocluido.

Un ejemplo se muestra en la Figura 4. 5, donde el valor de “V” para el caso de (a) se mantiene en un valor mayor al 96% en todos los puntos característicos, mientras que para (b), el valor de “V” asociado a la rodilla recae a un valor de 88% y para el caso de (c) el valor los puntos característicos recaen a un valor menor del 50% por lo que no se visibilizan en la imagen, sin embargo, *BlazePose* aún es capaz de identificar el punto asociado al tobillo en un 70%.

Figura 4. 5 Visualización de oclusiones parciales con estimación de la pose por BlazePose con imagen del CrowdPose

Tomando esto en cuenta se realizó la experimentación sobre los conjuntos de datos propios, generando con ayuda de *Open CV* oclusiones artificiales sobre los videos, para

posteriormente realizar la evaluación de los modelos de clasificación sobre el conjunto de datos con oclusiones parciales.

En la Figura 4. 6 se muestra los cuadros generados de manera aleatoria sobre los videos de entrenamiento. Para la generación de ellos se definió de manera manual los posibles tamaños para su aparición aleatoria.

Figura 4. 6 Oclusiones parciales generadas sobre videos de entrenamiento

Para la evaluación y entrenamiento de los modelos de clasificación sobre el conjunto de datos con oclusiones se consideró entrenar cada modelo a través de validación cruzada, esto con la finalidad de evitar el sobre ajuste que se presentó en las primeras pruebas de evaluación, donde los modelos alcanzaban hasta un 99% de exactitud, pero en el proceso de inferencia el rendimiento presentaba problemas.

En la Tabla 4. 7 se muestran los resultados de los 4 modelos de clasificación, así como sus tiempos de entrenamiento e inferencia sobre el conjunto de datos propios de 5,823 imágenes y 9 clases. Igualmente se destaca que los tiempos de entrenamiento para *random forest, ridge classifier* y *logistic regression* no pasan de los 0.5 segundos, mientras que el entrenamiento para *gradient boosting tiene un promedio de 120 segundos.*

Tabla 4. 7 resultados de entrenamiento de los modelos de clasificación con validación cruzada

Random forest	n_estimators = 10	Criterion {“gini”}		
Tiempo de entrenamiento	Tiempo de inferencia	Exactitud	Precisión	Sensibilidad
0.3361001	0.00997353	0.89022298	0.90459144	0.89022298
0.3380971	0.00897479	0.8593482	0.86942725	0.8593482
0.33011723	0.00997233	0.86106346	0.90623766	0.86106346
0.32164788	**0.00997472**	**0.92783505**	**0.93344537**	**0.92783505**
0.34607387	0.00897646	0.86597938	0.89903579	0.86597938
Logistic Regression				
Tiempo de entrenamiento	Tiempo de inferencia	Exactitud	Precisión	Sensibilidad
0.43537688	0.01894927	0.86523605	0.87323186	0.86523605
0.42053556	0.01894903	0.89356223	0.90080499	0.89356223
0.37610483	**0.01991248**	**0.90214592**	**0.91401244**	**0.90214592**
0.37151551	0.01994562	0.85137457	0.88462714	0.85137457
0.49602652	0.01894784	0.83247423	0.8648078	0.83247423
Ridge Classification	solver = auto			
Tiempo de entrenamiento	Tiempo de inferencia	Exactitud	Precisión	Sensibilidad
0.06688833	0.01795149	0.85236052	0.87940842	0.85236052
0.05186129	0.01816058	0.81287554	0.81162775	0.81287554
0.04886842	**0.01595974**	**0.85922747**	**0.88103544**	**0.85922747**
0.0498991	0.01791978	0.80756014	0.8243672	0.80756014
0.0588429	0.02194405	0.70017182	0.72085924	0.70017182
Gradient Boosting	n_estimators=50	criterion=friedman_mse		
Tiempo de entrenamiento	Tiempo de inferencia	Exactitud	Precisión	Sensibilidad
126.9989209	0.02194142	0.82403433	0.81440308	0.82403433
125.852963	**0.02194309**	**0.89785408**	**0.91090832**	**0.89785408**
129.342284	0.02293944	0.79141631	0.82444132	0.79141631
122.5245416	0.02194118	0.88745704	0.90536824	0.88745704
124.0477197	0.02393413	0.87457045	0.88309415	0.87457045

Con los resultados obtenidos en la Tabla 4. 7 se puede apreciar que el rendimiento para los clasificadores no disminuye drásticamente, se mantiene en un valor cercano al 90% de exactitud, esto principalmente por la robustez que brinda el modelo de *BlazePose* para la estimación de los puntos característicos aun con presencia de oclusiones.

El modelo de *random forest* mostró consistencia como el algoritmo con mejores resultados, logrando en este caso experimental hasta el 92.7% de exactitud y el menor tiempo de inferencia, de solo 0.009 segundos. También cabe mencionar que los parámetros de los modelos no variaban mucho el rendimiento general de los resultados, pero sí con respecto a su tiempo de inferencia y entrenamiento, principalmente en los modelos ensamblados de random forest y gradient boosting.

4.4. Resultados

Con relación al experimento 1 se determinó el uso de BlazePose para la integración final, esto debido a la robustes que tiene el modelo de estimación de la pose a auto oclusiones y oclusiones parciales. También por la incorporación del seguimiento de los puntos característicos del cuerpo, teniendo así mejores resultados para la estimación de la pose en secuencia de imágenes. Igualmente cabe recalcar que el modelo muestra una gran robustez a cambios en las condiciones de iluminación.

En el experimento 2 se determinó que los modelos de clasificación de la pose tienen un buen rendimiento para la problemática abordada aun con la presencia de auto oclusiones. Sin embargo, al momento de medir el rendimiento en el subconjunto de prueba los resultados mostraban ser inferiores y de igual manera durante el proceso de inferencia, por lo que en esta etapa se decidió incorporar la codificación de los puntos característicos en términos de distancias, logrando así una mejoría en los resultados presentados en la tablas 4.5 y 4.6 respectivamente.

Finalmente, el experimento 3 ayudó a determinar el modelo de clasificación de la pose con mejores características durante la evaluación de su exactitud y también considerando su tiempo de inferencia menor al de los demás modelos evaluados.

CAPÍTULO 5 CONCLUSIONES Y TRABAJO FUTURO

En esta sección se presentan los resultados obtenidos a través de la comparativa con los objetivos planteados en un principio para el complimiento del proyecto, así como una conclusión con respecto a las aportaciones de este trabajo y posibles trabajos futuros; de igual manera se presentan las actividades académicas realizadas en los diversos semestres.

5.1. Conclusiones

En este trabajo de investigación se diseñó y desarrolló un sistema de visión artificial que estima un conjunto de poses humanas previamente definidas a partir de un análisis con mayor peso a la identificación de poses con auto oclusiones y oclusiones parciales reportadas en la literatura, planteando así una metodología que hace uso de los modelos de estimación de la pose reportados y de modelos de clasificación basados en aprendizaje automático.

Los modelos de estimación de la pose reportados en la literatura van en aumento, esto debido al crecimiento en la investigación del área y a la alta demanda de aplicaciones basadas en estos modelos, por lo que hoy en día dichos modelos presentan gran robustez a problemas como rotación, traslación, cambio de dimensiones en la imagen, variabilidad en tonalidades de piel, e inclusive oclusiones. Por esto mismo, la implementación de estos modelos para abordar la estimación o clasificación de la pose como objetivo principal

facilita otras metodologías como el uso de descriptores basados en forma, uso de dispositivos inteligentes y procesamiento de la imagen completa mediante redes neuronales.

El modelo de estimación de la pose que mostró un mejor rendimiento basado en la tabla 4.1 y 4.2 fue *YoloPose*, sin embargo, *BlazePose* es capaz de realizar la estimación de la pose aun con la presencia de oclusiones en videos debido al seguimiento de los puntos característicos y también de ejecutarse en tiempo real debido a su arquitectura.

Por otra parte, los modelos de clasificación basados en aprendizaje automático demuestran un resultado mayor al 90 % de exactitud para realizar la tarea de clasificación de la pose a través de los puntos característicos del modelo de estimación de la pose. También se muestra que la codificación de dichos puntos a través del cálculo de distancia entre ellos permite una mayor generalización intraclase al momento de inferencia.

De manera general el tiempo de respuesta para la inferencia del modelo de clasificación es bastante aceptable con tiempos promedios de 0.01 segundo, esto con base en los resultados obtenidos en la tabla 4.7.

5.2. Objetivos

OBJETIVO GENERAL	
Diseñar un sistema de visión artificial que estime un conjunto de poses, a pesar de presentar oclusiones parciales, a partir de una imagen o secuencia de imágenes.	Para el complimiento del objetivo se diseñó y desarrollo un sistema de visión artificial que estima la pose humana y la clasifica en un conjunto de poses definidas bajo ciertos criterios de oclusión esto sobre una imagen o secuencia de imágenes.
OBJETIVOS ESPECIFICOS	
Revisar en el estado del arte las técnicas utilizadas para la estimación de la pose.	Se revisó un total de 60 artículos asociados a la estimación de la pose en sus diferentes técnicas, de igual manera se reportó un documento con 25

	de los modelos más mencionados y utilizados en la literatura.
Diseñar un sistema de visión artificial autónomo capaz de estimar la pose humana en un área controlada.	Se diseñó un sistema de visión artificial en Python que implementa un modelo de estimación de la pose humana para llevar a cabo la clasificación de la pose en un área controlada.
	Se realizó el análisis de 4 modelos de estimación de la pose disponibles en la literatura a través de la evaluación de ellos en dos conjuntos de datos, OcHuaman y ClowdPose
	Se realizó la avaluación de 4 modelos de aprendizaje automático para la clasificación de la pose humana a través de 3 métricas, precisión, sensibilidad y exactitud.

5.3. Aportaciones

La principal aportación de esta investigación radica en el diseño y desarrollo de un sistema de visión artificial que estima la pose humana de actividades no cotidianas con presencia de oclusiones parciales para posteriormente clasificarla en alguna pose previamente entrenada.

Algunas de las características del sistema son las siguientes:

Entrenamiento:

- ✓ Adquisición de video a través de cámara web
- ✓ Selección de video en formato .avi o .mp4
- ✓ Selección de modelos de estimación de la pose
- ✓ Recolección de puntos característicos

- ✓ Creación de archivo .txt con anotaciones para hasta 10 clases
- ✓ Entrenamiento de clasificación con 4 modelos de aprendizaje automático
- ✓ Serializado de modelo a formato *.pkl* para exportación

Inferencia:

- ✓ Adquisición de video a través de cámara web
- ✓ Selección de video en formato *.avi* o *.mp4*
- ✓ Selección de modelo de estimación de la pose
- ✓ Importación de modelo de clasificación serializado
- ✓ Inferencia para estimación de la pose y clasificación

Validación:

- ✓ Visualización gráfica de la matriz de confusión de cada modelo

De igual manera se creó un conjunto de datos, con 5,800 anotaciones de 9 poses realizadas por 3 personas con diferentes complexiones, algunas características son:

- ✓ Variabilidad de poses intraclase
- ✓ Poses con presencia de auto oclusiones
- ✓ 37 videos con duración de un minuto por clase
- ✓ Anotaciones de cada pose de ejemplo y su respectivo video de origen, así como su cuadro de imagen del video

5.4. Trabajos futuros

Cabe mencionar que el uso de la metodología abordada para el cumplimiento de los objetivo se puede implementar de manera efectiva a los modelos de reconocimiento de actividades humanas como trabajo futuro, esto porque en dicha área se puede clasificar las actividades en estáticas o dinámicas, o bien una combinación o transición entre ellas, por lo que no solo se podría realizar la clasificación de la pose humana sino también

implementar el reconocimiento de actividades realizadas por la persona o de igual manera su interacción con objetos.

Por lo que se puede trabajar en un futuro en:

- Identificar la actividad humana a través de modelos secuenciales haciendo uso de la clasificación de la pose humana.
- Realizar el reconocimiento de objetos y su interacción entre la persona de estudio por medio de la estimación y clasificación de la pose.

5.5. Actividades académicas

En esta sección se presentan los productos académicos presentados a lo largo de la maestría en orden cronológico.

En la Figura 5. 1 se presenta el reconocimiento por la presentación del poster “Estimación de la pose humana en secuencia de imágenes” presentado en la escuela de inteligencia computacional y robótica en la UTEZ en agosto del 2022.

Figura 5. 1 Reconocimiento por exposición de poster en la escuela de inteligencia computacional y robótica realizada en la UTEZ

En la Figura 5. 2 se presenta el reconocimiento por la presentación del artículo “estimación de la pose humana en secuencia de imágenes” en la novena jornada ciencia y tecnología del CENIDET en noviembre de 2022.

EDUCACIÓN
SECRETARÍA DE EDUCACIÓN PÚBLICA

TECNOLOGICO NACIONAL DE MÉXICO

EL TECNOLÓGICO NACIONAL DE MÉXICO
A TRAVÉS DEL CENTRO NACIONAL DE INVESTIGACIÓN Y DESARROLLO TECNOLÓGICO

OTORGA EL PRESENTE

RECONOCIMIENTO

A

LUIS PAUL ARRIAGA SANTOS
CENTRO NACIONAL DE INVESTIGACIÓN Y DESARROLLO TECNOLÓGICO

POR LA PRESENTACIÓN DEL ARTICULO:
ESTIMACIÓN DE LA POSICIÓN DEL CUERPO HUMANO EN SECUENCIA DE IMÁGENES
EN EL MARCO DE LA 9ª JORNADA DE CIENCIA Y TECNOLOGÍA APLICADA, CELEBRADA
DEL 16 AL 18 DE NOVIEMBRE DE 2022, EN EL TECNM/CENIDET

CUERNAVACA, MORELOS, 16-18 DE NOVIEMBRE DE 2022

AL0107622
http://constancias.cenidet.tecnm.mx

DRA. YESICA IMELDA SAAVEDRA BENÍTEZ
DIRECTORA DEL CENTRO NACIONAL DE INVESTIGACIÓN Y DESARROLLO TECNOLÓGICO

Sello Digital:

cenidet
Centro Nacional de Investigación y Desarrollo Tecnológico

JORNADA DE CIENCIA & TECNOLOGÍA APLICADA

Figura 5. 2 Reconocimiento por presentación de artículo en la novena jornada ciencia y tecnología aplicada

En la Figura 5. 3 se presenta el reconocimiento por la presentación de proyecto de investigación de la línea de inteligencia artificial en el marco del simposio internacional de ingeniería en sistemas computacionales en el instituto tecnológico de Cuautla en marzo del 2023.


EDUCACIÓN
SECRETARÍA DE EDUCACIÓN PÚBLICA

TECNOLÓGICO NACIONAL DE MÉXICO

EL TECNOLÓGICO NACIONAL DE MÉXICO
A TRAVÉS DEL INSTITUTO TECNOLÓGICO DE CUAUTLA

OTORGA EL PRESENTE

RECONOCIMIENTO

A

LUIS PAUL ARRIAGA SANTOS

POR SU VALIOSA PARTICIPACIÓN COMO PONENTE EN LA CONFERENCIA:

"PROYECTOS DE INVESTIGACIÓN DE LA LÍNEA DE INTELIGENCIA ARTIFICIAL".

IMPARTIDA EL DÍA 29 DE MARZO DE 2023, EN EL MARCO DEL SIMPOSIO INTERNACIONAL DE INGENIERÍA EN SISTEMAS COMPUTACIONALES – SISTEMAS TRANSVERSALES.

H. H. Cuautla, Morelos a 29 de marzo de 2023

PORFIRIO ROBERTO NÁJERA MEDINA
DIRECTOR DEL INSTITUTO TECNOLÓGICO DE CUAUTLA

Figura 5. 3 Reconocimiento por presentación en el marco del simposio internacional de ingeniería en sistemas computacionales en el instituto tecnológico de Cuautla

Como último en la Figura 5. 4 se presenta el reconocimiento por la presentación del artículo “Clasificación de la pose huma a través de puntos característicos” presentado en la décima jornada de ciencia y tecnología aplicada en abril del 2023.


EDUCACIÓN
SECRETARÍA DE EDUCACIÓN PÚBLICA

TECNOLÓGICO NACIONAL DE MÉXICO

EL TECNOLÓGICO NACIONAL DE MÉXICO
A TRAVÉS DEL CENTRO NACIONAL DE INVESTIGACIÓN
Y DESARROLLO TECNOLÓGICO

OTORGA EL PRESENTE

RECONOCIMIENTO

A

LUIS PAUL ARRIAGA SANTOS

CENIDET

POR LA EXPOSICIÓN DE MANERA PRESENCIAL DEL ARTICULO:
CLASIFICACIÓN DE LA POSE DEL CUERPO HUMANO A TRAVÉS DE PUNTOS CARACTERÍSTICOS
EN EL MARCO DE LA X JORNADA DE CIENCIA Y TECNOLOGÍA APLICADA, CELEBRADA DEL 26 AL 28
DE ABRIL DE 2023, EN EL TECNM/CENIDET

CUERNAVACA, MORELOS, 28 DE ABRIL DE 2023

JCYTA054723

DR. ARTURO ERNESTO MARES GARDEA
DIRECTOR DEL CENTRO NACIONAL DE INVESTIGACIÓN
Y DESARROLLO TECNOLÓGICO

Sello Digital:

cenidet
Centro Nacional de Investigación y Desarrollo Tecnológico

X JORNADA DE CIENCIA Y TECNOLOGÍA APLICADA

100% PLÁSTICO

Figura 5. 4 Reconocimiento por presentación de artículo científico en la décima jornada ciencia y tecnología

6. Referencias

Alshari, H.H., Saleh, A.Y. and Odabaş, A. (2021) 'Comparison of Gradient Boosting Decision Tree Algorithms for CPU Performance', *Journal of Institue Of Science and Technology*, 37(1).

Ansarian, A. and Amer, M.A. (2021a) 'REALISTIC AUGMENTATION FOR EFFECTIVE 2D HUMAN POSE ESTIMATION UNDER OCCLUSION', in *Proceedings - International Conference on Image Processing, ICIP*. Available at: https://doi.org/10.1109/ICIP42928.2021.9506392.

Ansarian, A. and Amer, M.A. (2021b) 'REALISTIC AUGMENTATION FOR EFFECTIVE 2D HUMAN POSE ESTIMATION UNDER OCCLUSION', in *Proceedings - International Conference on Image Processing, ICIP*. Available at: https://doi.org/10.1109/ICIP42928.2021.9506392.

Bazarevsky, V. and Grishchenko, I. (2020a) *Google AI Blog: On-device, Real-time Body Pose Tracking with MediaPipe BlazePose*, *Google AI Blog*.

Bazarevsky, V. and Grishchenko, I. (2020b) *Google AI Blog: On-device, Real-time Body Pose Tracking with MediaPipe BlazePose*, *Google AI Blog*.

Belagiannis, V. *et al.* (2016) '3D Pictorial Structures Revisited: Multiple Human Pose Estimation', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(10). Available at: https://doi.org/10.1109/TPAMI.2015.2509986.

Bharadwaj, Prakash, K.B. and Kanagachidambaresan, G.R. (2021) 'Pattern Recognition and Machine Learning', in *EAI/Springer Innovations in Communication and Computing*. Available at: https://doi.org/10.1007/978-3-030-57077-4_11.

Breiman, L. (2001) 'Random forests', *Machine Learning*, 45(1). Available at: https://doi.org/10.1023/A:1010933404324.

Cao, Z. *et al.* (2017a) 'Realtime multi-person 2D pose estimation using part affinity fields', in *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*. Available at: https://doi.org/10.1109/CVPR.2017.143.

Cao, Z. *et al.* (2017b) 'Realtime multi-person 2D pose estimation using part affinity fields', in *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*. Available at: https://doi.org/10.1109/CVPR.2017.143.

Cao, Z. *et al.* (2021) 'OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(1). Available at: https://doi.org/10.1109/TPAMI.2019.2929257.

Chen, Y., Tian, Y. and He, M. (2020) 'Monocular human pose estimation: A survey of deep learning-based methods', *Computer Vision and Image Understanding*, 192. Available at: https://doi.org/10.1016/j.cviu.2019.102897.

Dang, Q. *et al.* (2019a) 'Deep learning based 2D human pose estimation: A survey', *Tsinghua Science and Technology*, 24(6). Available at: https://doi.org/10.26599/TST.2018.9010100.

Dang, Q. *et al.* (2019b) 'Deep learning based 2D human pose estimation: A survey', *Tsinghua Science and Technology*, 24(6). Available at: https://doi.org/10.26599/TST.2018.9010100.

Debapriya M, Soyeb N and Manu M (2022a) 'YOLO-Pose - Enhancing YOLO for Multi Person Pose Estimation Using Object Keypoint Similarity Loss', *DeepAI* [Preprint].

Debapriya M, Soyeb N and Manu M (2022b) 'YOLO-Pose - Enhancing YOLO for Multi Person Pose Estimation Using Object Keypoint Similarity Loss', *DeepAI* [Preprint].

Deepa, N. *et al.* (2021) 'An AI-based intelligent system for healthcare analysis using Ridge-Adaline Stochastic Gradient Descent Classifier', *Journal of Supercomputing*, 77(2). Available at: https://doi.org/10.1007/s11227-020-03347-2.

Fürst, M. *et al.* (2020) 'HPERL: 3D human pose estimation from RgB and LiDAR', in *Proceedings - International Conference on Pattern Recognition*. Available at: https://doi.org/10.1109/ICPR48806.2021.9412785.

Ben Gamra, M. and Akhloufi, M.A. (2021) 'A review of deep learning techniques for 2D and 3D human pose estimation', *Image and Vision Computing*. Available at: https://doi.org/10.1016/j.imavis.2021.104282.

Geng, Z. *et al.* (2021a) 'Bottom-up human pose estimation via disentangled keypoint regression', in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. Available at: https://doi.org/10.1109/CVPR46437.2021.01444.

Geng, Z. *et al.* (2021b) 'Bottom-up human pose estimation via disentangled keypoint regression', in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. Available at: https://doi.org/10.1109/CVPR46437.2021.01444.

Ghafoor, M. and Mahmood, A. (2022) 'Quantification of Occlusion Handling Capability of 3D Human Pose Estimation Framework', *IEEE Transactions on Multimedia* [Preprint]. Available at: https://doi.org/10.1109/TMM.2022.3158068.

Huang, X., Huang, J. and Tang, Z. (2021) '3D Human Pose Estimation with Spatial Structure Information', *IEEE Access*, 9. Available at: https://doi.org/10.1109/ACCESS.2021.3062426.

Hwang, J., Yang, J. and Kwak, N. (2020) 'Exploring rare pose in human pose estimation', *IEEE Access*, 8. Available at: https://doi.org/10.1109/ACCESS.2020.3033531.

Jorge Alfredo Saldaña García (2007) *Caracterización de Imágenes en Movimiento: Correr y Caminar*. CENIDET.

Kato, N., Honda, H. and Uchida, Y. (2020) 'Leveraging Temporal Joint Depths for Improving 3D Human Pose Estimation in Video', in *2020 IEEE 9th Global Conference on Consumer Electronics, GCCE 2020*. Available at: https://doi.org/10.1109/GCCE50665.2020.9292014.

Khan, N.U. and Wan, W. (2018) 'A Review of Human Pose Estimation from Single Image', in *ICALIP 2018 - 6th International Conference on Audio, Language and Image Processing*. Available at: https://doi.org/10.1109/ICALIP.2018.8455796.

Kherwa, P. *et al.* (2021) 'Articulated Human Pose Estimation Using Greedy Approach', in. Available at: https://doi.org/10.5772/intechopen.99354.

Kilbas, I., Gribanov, D. and Paringer, R. (2022a) 'A neural network based algorithm for classification of sets of human body keypoints', in *2022 8th International Conference on Information Technology and Nanotechnology, ITNT 2022*. Institute of Electrical and Electronics Engineers Inc. Available at: https://doi.org/10.1109/ITNT55410.2022.9848751.

Kilbas, I., Gribanov, D. and Paringer, R. (2022b) 'A neural network based algorithm for classification of sets of human body keypoints', in *2022 8th International Conference on Information Technology and Nanotechnology, ITNT 2022*. Institute of Electrical and Electronics Engineers Inc. Available at: https://doi.org/10.1109/ITNT55410.2022.9848751.

Li, J. *et al.* (2019a) 'Crowdpose: Efficient crowded scenes pose estimation and a new benchmark', in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. Available at: https://doi.org/10.1109/CVPR.2019.01112.

Li, J. *et al.* (2019b) 'Crowdpose: Efficient crowded scenes pose estimation and a new benchmark', in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. Available at: https://doi.org/10.1109/CVPR.2019.01112.

María Luisa Dávila García (2009) *Caracterización automática de una muestra de acciones en vídeo*. CENIDET.

Miao, J., Wu, Y. and Yang, Y. (2021) 'Identifying Visible Parts via Pose Estimation for Occluded Person Re-Identification', *IEEE Transactions on Neural Networks and Learning Systems* [Preprint]. Available at: https://doi.org/10.1109/TNNLS.2021.3059515.

Mroz, S. *et al.* (2021) 'Comparing the Quality of Human Pose Estimation with BlazePose or OpenPose', in *BioSMART 2021 - Proceedings: 4th International Conference on Bio-Engineering for Smart Technologies*. Available at: https://doi.org/10.1109/BioSMART54244.2021.9677850.

Munea, T.L. *et al.* (2020) 'The Progress of Human Pose Estimation: A Survey and Taxonomy of Models Applied in 2D Human Pose Estimation', *IEEE Access*, 8. Available at: https://doi.org/10.1109/ACCESS.2020.3010248.

Newell, A., Yang, K. and Deng, J. (2016) 'Stacked hourglass networks for human pose estimation', in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Available at: https://doi.org/10.1007/978-3-319-46484-8_29.

Park, S., Ji, M. and Chun, J. (2018) '2D human pose estimation based on object detection using RGB-D information', *KSII Transactions on Internet and Information Systems*, 12(2). Available at: https://doi.org/10.3837/tiis.2018.02.015.

Rafael Alcantar Juárez (2013) *Caracterización Visual de Movimientos "Sospechosos" de Personas en Estacionamientos*. CENIDET.

Sabouri, S. *et al.* (2020) 'Logistic regression', in *Basic Quantitative Research Methods for Urban Planners*. Available at: https://doi.org/10.4324/9780429325021-13.

Sun, K. *et al.* (2019) 'Deep high-resolution representation learning for human pose estimation', in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. Available at: https://doi.org/10.1109/CVPR.2019.00584.

Tang, H., Wang, Q. and Chen, H. (2019a) 'Research on 3D human pose estimation using RGBD camera', in *ICEIEC 2019 - Proceedings of 2019 IEEE 9th International Conference on Electronics Information and Emergency Communication*. Available at: https://doi.org/10.1109/ICEIEC.2019.8784591.

Tang, H., Wang, Q. and Chen, H. (2019b) 'Research on 3D human pose estimation using RGBD camera', in *ICEIEC 2019 - Proceedings of 2019 IEEE 9th International Conference on Electronics Information and Emergency Communication*. Available at: https://doi.org/10.1109/ICEIEC.2019.8784591.

Wang, C. *et al.* (2022) 'Low-resolution human pose estimation', *Pattern Recognition*, 126. Available at: https://doi.org/10.1016/j.patcog.2022.108579.

Wang, J. *et al.* (2021) 'Deep High-Resolution Representation Learning for Visual Recognition', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(10). Available at: https://doi.org/10.1109/TPAMI.2020.2983686.

Yang, S. *et al.* (2020) 'TransPose: Keypoint Localization via Transformer'. Available at: http://arxiv.org/abs/2012.14214.

Yu, X. *et al.* (2022) 'Overlapped Human Pose Estimation using Non-Maximum Suppression based on Shape Similarity', in. Available at: https://doi.org/10.1109/cac53003.2021.9728073.

Zhang, S.H. *et al.* (2019a) 'Pose2Seg: Detection free human instance segmentation', in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. Available at: https://doi.org/10.1109/CVPR.2019.00098.

Zhang, S.H. *et al.* (2019b) 'Pose2Seg: Detection free human instance segmentation', in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. Available at: https://doi.org/10.1109/CVPR.2019.00098.

ANEXOS

ANEXO A. - Configuración de conjunto de datos Crowdpose y OcHuaman

A través de MMPose se han establecido las configuraciones utilizadas para la validación de los modelos de estimación de la pose compatibles desde MMPose, tales como el subconjunto de imágenes seleccionado a través del *pipeline* de validación.

```
# pipelines
val_pipeline = [
    dict(type='LoadImage'),
    dict(type='GetBBoxCenterScale'),
    dict(type='Topdown', input_size=codec['480x640’]),
    dict(type='PackPoseInputs')
]
```

De esta manera se definió el proceso de validación, de igual manera se estableció el *dataloader* para el caso de *Crowdpose* bajo

```
# dataloader
val_dataloader = dict(
    batch_size=32,
    num_workers=2,
    persistent_workers=True,
    drop_last=False,
    sampler=dict(type='DefaultSampler', shuffle=False, round_up=False),
    dataset=dict(
        type=dataset_type,
        sub_data_size=200
        data_root=data_root,
        data_mode=data_mode,
        ann_file='D:/ClowdPose/Anotaciones/crowdpose_val.json',
        bbox_file='data/ ClowdPose
        /annotations/det_for_crowdpose_tesy_0.1_0.5.json/',
        data_prefix=dict(img='1000'),
        val_mode=True,
        pipeline=val_pipeline,
    ))
test_dataloader = val_dataloader
```

La estructura general para el conjunto de datos Crowdpose se establece en la Figura A.1

```
mmpose
├── mmpose
├── docs
├── tests
├── tools
├── configs
`── data
    |── crowdpose
        |-- annotations
        |   |-- mmpose_crowdpose_train.json
        |   |-- mmpose_crowdpose_val.json
        |   |-- mmpose_crowdpose_trainval.json
        |   |-- mmpose_crowdpose_test.json
        |   |-- det_for_crowd_test_0.1_0.5.json
        |-- images
            |-- 100000.jpg
            |-- 100001.jpg
            |-- 100002.jpg
            |-- ...
```

Figura A.1 Estructura de *dataset* para *CrowdPose*

De igual manera se estableció la carga del conjunto de datos para OcHuman siguiendo la misma estructura para el *pipeline* y el *dataloader* y la estructura de datos mencionada en MMPose.

```
# pipelines
val_pipeline = [
    dict(type='LoadImage'),
    dict(type='GetBBoxCenterScale'),
    dict(type='Topdown', input_size=codec['480x640’]),
    dict(type='PackPoseInputs')
]
```

```
# dataloader
val_dataloader = dict(
    batch_size=32,
    num_workers=2,
    persistent_workers=True,
    drop_last=False,
    sampler=dict(type='DefaultSampler', shuffle=False, round_up=False),
    dataset=dict(
        type=dataset_type,
        sub_data_size=200
```

```
        data_root=data_root,
        data_mode=data_mode,
        ann_file='D:/OcHuaman/Validacion/ochumanjson',
        data_prefix=dict(img='1000'),
        test_mode=True,
        pipeline=val_pipeline,
    ))
test_dataloader = val_dataloader
```

La estructura general para el conjunto de datos OcHuman se establece enla Figura A.2

```
mmpose
├── mmpose
├── docs
├── tests
├── tools
├── configs
`── data
    |── ochuman
        |-- annotations
        |   |-- ochuman_coco_format_val_range_0.00_1.00.json
        |   |-- ochuman_coco_format_test_range_0.00_1.00.json
        |-- images
            |-- 000001.jpg
            |-- 000002.jpg
            |-- 000003.jpg
            |-- ...
```

Figura A.2 Estructura de dataset para OcHuman