



ugr | Universidad
de **Granada**

TRABAJO FIN DE GRADO
INGENIERÍA INFORMÁTICA

**Estimación de la distancia cámara-sujeto
en fotografías faciales mediante técnicas
de aprendizaje profundo**

Autor
Iván Salinas López

Directores
Enrique Bermejo Nievas
Pablo Mesejo Santiago



ESCUELA TÉCNICA SUPERIOR DE INGENIERÍAS INFORMÁTICA Y DE
TELECOMUNICACIÓN

—
Granada, Junio de 2024

Estimación de distancia cámara-sujeto en fotografías faciales usando aprendizaje profundo

Iván Salinas López

Palabras clave: palabra_clave1, palabra_clave2, palabra_clave3,

Resumen

Poner aquí el resumen.

Camera-subject distance estimation in facial photographs using deep learning

Iván Salinas López

Keywords: Keyword1, Keyword2, Keyword3,

Abstract

Write here the abstract in English.

Yo, **Iván Salinas López**, alumno de la titulación Grado en Ingeniería Informática de la **Escuela Técnica Superior de Ingenierías Informática y de Telecomunicación de la Universidad de Granada**, con DNI 78026145W, autorizo la ubicación de la siguiente copia de mi Trabajo Fin de Grado en la biblioteca del centro para que pueda ser consultada por las personas que lo deseen.

Fdo: Iván Salinas López

Granada a X de Junio de 2024

D. **Enrique Bermejo Nievas**, Investigador Senior en Panacea Cooperative Research y miembro del Instituto Andaluz Interuniversitario en Ciencia de Datos e Inteligencia Computacional.

D. **Pablo Mesejo Santiago**, Profesor del Área de Ciencias de la Computación e Inteligencia Artificial del Departamento Ciencias de la Computación e Inteligencia Artificial de la Universidad de Granada.

Informan:

Que el presente trabajo, titulado *Estimación de la distancia cámara-sujeto en fotografías faciales mediante técnicas de aprendizaje profundo*, ha sido realizado bajo su supervisión por **Iván Salinas López**, y autorizamos la defensa de dicho trabajo ante el tribunal que corresponda.

Y para que conste, expiden y firman el presente informe en Granada a X de Junio de 2024.

Los directores:

Enrique Bermejo Nievas

Pablo Mesejo Santiago

Agradecimientos

Poner aquí agradecimientos...

Índice general

1. Introducción	1
1.1. Definición del problema	1
1.2. Motivación	4
1.3. Objetivos	5
1.4. Planificación del proyecto	5
2. Fundamentos teóricos	9
2.1. Aprendizaje automático	9
2.2. Aprendizaje profundo	10
2.2.1. Redes neuronales artificiales	10
2.2.2. Redes neuronales convolucionales	12
2.2.3. Transferencia de aprendizaje	16
2.2.4. Regularización	16
2.3. Parámetros de la cámara y perspectiva	17
3. Estado del Arte	23
3.1. Primeros enfoques	23
3.2. MediaPipe Iris	25
3.3. PerspectiveX	25
3.4. FacialSCDnet	26
4. Materiales y métodos	29
4.1. Materiales	29
4.1.1. Modelos 3D	29
4.1.2. Procesamiento del conjunto de datos	31
4.2. Métodos	32
4.2.1. FacialSCDnet+	32

Índice de figuras

1.1.	Número de publicaciones, en Scopus, relacionadas con imágenes faciales en los últimos 45 años.	1
1.2.	Efectos de la distorsión de perspectiva en fotografías faciales realizadas a diferentes distancias: 0.3 m, 0.6 m y 1.5 m respectivamente.	2
1.3.	Diagrama del proceso de estimación automática de la SCD en este proyecto.	5
2.1.	Esquema de una red neuronal [24].	10
2.2.	Modelo neuronal para una neurona k [20].	11
2.3.	Ejemplo de CNN [37].	12
2.4.	Ejemplo de convolución en CNN [48].	13
2.5.	Funciones de activación comúnmente aplicadas en CNN [48].	14
2.6.	Tipos de <i>pooling</i> comúnmente utilizados en CNN [48]. . . .	15
2.7.	Zona de infraajuste y sobreajuste durante el entrenamiento [48].	18
2.8.	Relación entre punto nodal y longitud focal [18]	18
2.9.	Tamaños del sensor expresados según el factor de recorte [6] .	19
2.10.	Ejemplo de longitud focal equivalente según el tamaño del sensor [41].	19
2.11.	Distancia desde la cámara al sujeto.	20
2.12.	Efecto de la distorsión conforme se acerca la cámara al objeto [13].	20
3.1.	Número de publicaciones, en Scopus, relacionadas con la estimación de la distancia en fotografías faciales en función del año de publicación	24
4.1.	HeadSpace 3D	30
4.2.	H3DS-net	30
4.3.	HuMMAn	30
4.4.	People Snapshot	31
4.5.	Render People	31

Índice de cuadros

1.1.	Planificación inicial del proyecto	6
1.2.	7

Capítulo 1

Introducción

1.1. Definición del problema

En la era digital actual, las imágenes faciales han adquirido una relevancia significativa (véase Figura 1.1), dado su amplio uso en aplicaciones multimedia, plataformas de redes sociales, sistemas de vigilancia/seguridad, así como en investigaciones criminales y forenses. Esta expansión se debe en gran medida al continuo desarrollo tecnológico, que ha mejorado tanto la calidad como la ubicuidad de las fotografías faciales, permitiendo su presencia en una variedad cada vez mayor de contextos y aplicaciones.

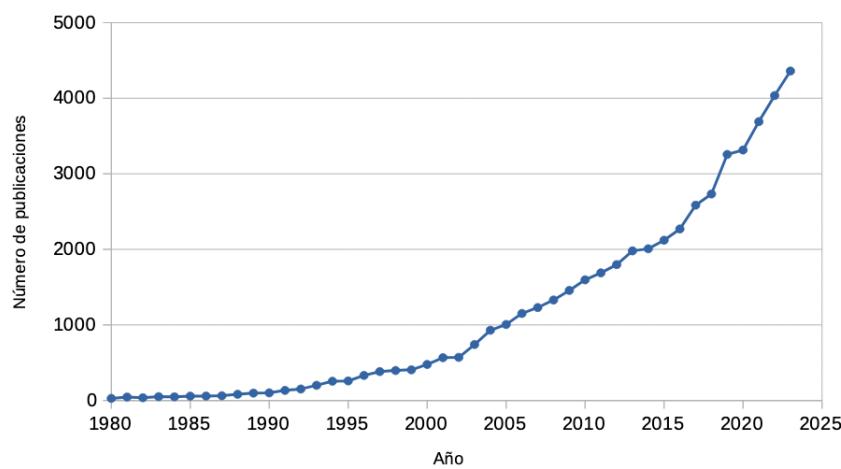


Figura 1.1: Número de publicaciones, en Scopus, relacionadas con imágenes faciales en los últimos 45 años.

En este contexto, es importante resaltar el papel que tienen las imágenes

faciales en campos como la biometría para la verificación de identidad, así como en la seguridad nacional, donde se pueden utilizar para la identificación facial. Para garantizar el correcto desempeño de estas aplicaciones, se debe tener en cuenta la calidad de las imágenes y todos los factores que afectan a la escena fotográfica. Por ello, existen numerosas herramientas y técnicas dirigidas a la extracción de metadatos, la detección facial o la estimación de la pose [50], todas ellas fundamentales para asegurar la fiabilidad y precisión de los sistemas de identificación facial.

En el ámbito forense, una de las técnicas más empleadas en la identificación facial es la comparación facial forense (CFF) [15]. Esta técnica, llevada a cabo por expertos manualmente o con la ayuda de sistemas automáticos, consiste en identificar similitudes y diferencias entre dos o más imágenes con el objetivo de determinar si representan a la misma persona. Para que este análisis sea confiable y concluyente, las imágenes faciales deben estar en unas condiciones adecuadas. Aspectos como la calidad, la resolución, el enfoque o la iluminación deben cumplir unos requisitos mínimos. Además, es importante que las características de la escena, como el ángulo de la cámara, la posición de la cabeza y la expresión facial, no varíen significativamente, con el fin de asegurar la similitud entre las imágenes y permitir una comparación precisa entre pares [14, 39].

Uno de los factores más importantes a tener en cuenta en las fotografías faciales es la distorsión de perspectiva, la cual puede provocar deformaciones en los rasgos faciales, como en las orejas, la nariz o la forma general del rostro, especialmente cuando la cámara está muy cerca del sujeto al momento de tomar la fotografía [32] (ver Figura 1.2). Esta alteración en la perspectiva repercute negativamente tanto en los sistemas de reconocimiento facial como en la CFF, complicando el análisis visual al alterar como se perciben ciertos rasgos.



Figura 1.2: Efectos de la distorsión de perspectiva en fotografías faciales realizadas a diferentes distancias: 0.3 m, 0.6 m y 1.5 m respectivamente.

La distorsión de perspectiva está estrechamente relacionada con la distancia cámara-sujeto (*subject-to-camera distance*, SCD en adelante). Esta relación es de decrecimiento logarítmico, lo que significa que a distancias cor-

tas se produce una mayor distorsión, la cual va disminuyendo a medida que la distancia entre la cámara y el sujeto aumenta [40]. Conocer la SCD en fotografías faciales permite cuantificar la cantidad de distorsión presente en una imagen, así como las diferencias en la distorsión entre dos pares de imágenes. Esta información puede ser determinante a la hora de evaluar la identidad de un individuo al aplicar la técnica de CFF. Además, conocer la SCD facilita el desarrollo de técnicas que permitan corregir con precisión dicha distorsión [47].

La SCD, a diferencia de otros parámetros de la cámara como la longitud focal o el tamaño del sensor, no puede obtenerse directamente desde los metadatos de la fotografía [43]. Por tanto, se necesita un método preciso para su estimación.

En los últimos años se han utilizado varios métodos que combinan técnicas manuales y automatizadas basadas en puntos de referencia o en características anatómicas de la cara[16, 10]. Sin embargo, no se han obtenido resultados favorables debido a la dificultad para obtener estimaciones precisas en largas distancias por la diversa fisionomía de la cara, y a los problemas relacionados con los parámetros de la cámara, como el recorte de imágenes o la combinación de diferentes longitudes focales en el mismo conjunto de datos.

Hasta la fecha, uno de los métodos completamente automatizados para estimar la SCD en fotografías faciales es conocido como FacialSCDnet [4]. Este método emplea una arquitectura basada en aprendizaje profundo para procesar las imágenes faciales y calcular su SCD correspondiente. Sin embargo, FacialSCDnet presenta ciertas limitaciones. En primer lugar, el conjunto de datos utilizado es limitado, debido al número reducido de individuos y al hecho de que solo incluye adultos. Además, el conjunto de datos utilizado solo incluye modelos faciales. Estos factores aumentan el riesgo de sesgo en el modelo, lo que puede afectar a su capacidad para realizar estimaciones precisas en poblaciones más diversas.

Considerando todos estos aspectos, el presente Trabajo de Fin de Grado (TFG) pretende mejorar el método actual del estado del arte en la estimación automática de la distancia cámara-sujeto en fotografías faciales. Para ello, partimos de FacialSCDnet como una prueba de concepto sólida, reconociendo sus ventajas pero también identificando sus limitaciones inherentes. Nos centraremos en incorporar mejoras significativas que permitan solventar estas limitaciones con el objetivo de elevar el rendimiento y la precisión del sistema.

1.2. Motivación

En el campo del aprendizaje profundo, es común encontrar sesgos en los datos que pueden afectar la precisión y la capacidad de generalización de las soluciones. Estos sesgos pueden surgir debido a varios factores, como la falta de diversidad o el ruido excesivo en el conjunto de datos utilizado para entrenar los modelos. Este fenómeno también se observa en el método FacialSCDnet. Por lo tanto, para mejorar la calidad del conjunto de datos, sería beneficioso incorporar una mayor diversidad de sujetos en términos de edad, sexo biológico, ancestría, expresiones faciales, condiciones de iluminación y fondos, así como la inclusión de modelos tanto faciales como de cuerpo completo.

Una estrategia para abordar este desafío consiste en integrar múltiples bases de datos con el objetivo de construir un conjunto de datos más completo y diverso. Esta mejora contribuiría a una mejor capacidad de generalización y adaptación del modelo, al mitigar los sesgos inherentes a los datos.

Por otro lado, obtener conjuntos de datos reales de alta calidad no siempre es una tarea sencilla. La recopilación y etiquetado de datos pueden resultar costosos y requerir mucho tiempo. En muchos casos, los conjuntos de datos reales disponibles pueden ser limitados en términos de tamaño y diversidad, como ocurre con FacialSCDnet. Una solución ampliamente utilizada consiste en emplear conjuntos de datos sintéticos en lugar de conjuntos de datos reales [42, 19, 46]. El uso de conjuntos de datos sintéticos puede ayudar a reducir costos y tiempo de recopilación de datos, manteniendo o mejorando el rendimiento de los modelos.

Dentro del contexto de FacialSCDnet, una restricción clave reside en la necesidad de conocer la longitud focal y desarrollar un modelo en función de esta. Utilizar imágenes sintéticas, facilita considerablemente la creación de conjuntos de datos según la focal deseada. Al emplear datos sintéticos, se tiene un mayor control sobre los parámetros de generación, lo que permite ajustar la focal de manera precisa y reproducible.

Con el propósito de aumentar la calidad de los conjuntos de datos empleados en el aprendizaje profundo, particularmente en el contexto de FacialSCDnet, este trabajo se enfoca en reducir sesgos al integrar una mayor diversidad de sujetos y condiciones, además de optimizar los recursos temporales y financieros mediante el uso de conjuntos de datos sintéticos. Este enfoque combinado busca mejorar la capacidad actual de generalización y adaptación de los modelos de FacialSCDnet.

1.3. Objetivos

El objetivo general de este TFG consiste en desarrollar un mejor modelo de aprendizaje profundo para mejorar la estimación de la distancia cámara-sujeto en fotografías faciales. Para el desarrollo del proyecto, dividiremos el objetivo general en una serie de objetivos parciales:

1. Realizar un análisis exhaustivo del estado del arte y de las bases de datos de modelos faciales y humanos 3D.
2. Desarrollar un protocolo de estandarización y generación de imágenes sintéticas fotorrealistas.
3. Realizar un estudio comparativo y analizar la viabilidad de la nueva aproximación propuesta.
4. Explorar el uso de arquitecturas y tecnologías alternativas que permitan mejorar el rendimiento y/o los resultados del método original.

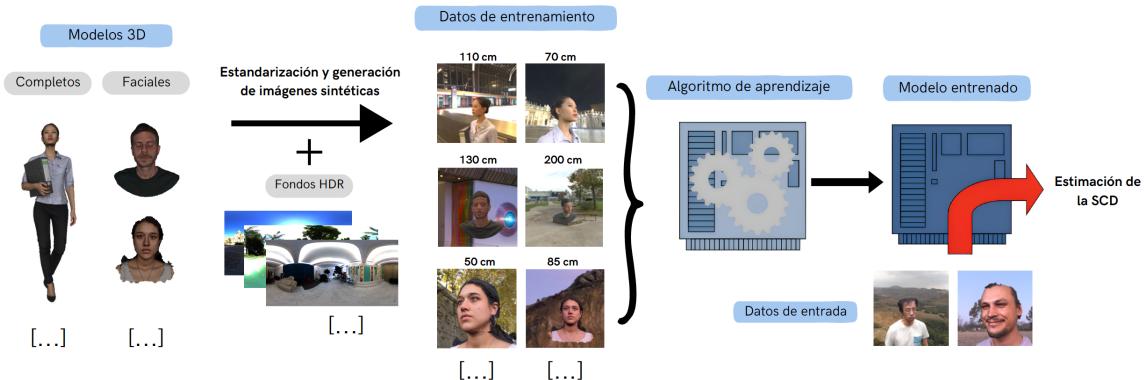


Figura 1.3: Diagrama del proceso de estimación automática de la SCD en este proyecto.

En este contexto, es relevante citar el trabajo de referencia [4], el cual proporciona una base sólida para el desarrollo del TFG.

1.4. Planificación del proyecto

Para abordar el desarrollo de este proyecto, es esencial considerar que el TFG tiene asignados 12 créditos ECTS, lo que equivale a aproximadamente 300 horas de trabajo. Dada la distribución temporal del segundo cuatrimestre, con unas 20 semanas disponibles, se estima que se requerirá dedicar al TFG unas 20 horas semanales, equivalentes a 4 horas diarias durante 5 días

a la semana. Se reservan así 4 semanas como margen para posibles retrasos o imprevistos que puedan surgir durante el desarrollo del proyecto.

En cuanto a la metodología de desarrollo, se ha optado por seguir un enfoque basado en el ciclo de vida en cascada [33], aunque con una variante que permite retroalimentación. Aunque el proyecto presenta requisitos y objetivos claros, se reconoce la posibilidad de ajustes menores durante su desarrollo, especialmente a medida que se obtenga más información sobre el problema y los métodos. Esta flexibilidad se considera crucial para adaptarse a posibles cambios en el contexto o los requisitos del proyecto.

Las fases del ciclo de vida del proyecto son las siguientes:

A continuación se describen las fases del ciclo de vida del proyecto:

- Análisis de Requisitos: Consiste en las reuniones iniciales con los clientes, en este caso los directores del TFG. Se realiza un análisis del problema y un estudio detallado de la bibliografía existente
- Diseño: Consiste en la exploración y selección de los métodos apropiados así como de los conjuntos de datos basados en el análisis previo, tanto para la resolución como para la validación de la solución propuesta. Además, se llevarán a cabo pruebas preliminares y se elaborará el diseño del software experimental.
- Implementación: Consiste en la adaptación del código de los modelos investigados, la implementación de nuevas funcionalidades y la generación de un conjunto de datos sintético junto con su posterior preprocesado.
- Pruebas: Consiste en la realización de diversos experimentos para validar el funcionamiento del software desarrollado, utilizando los modelos y datos previamente definidos.

Tarea	Semanas - Horas	Febrero				Marzo				Abril				Mayo				Junio			
		5	12	19	26	4	11	18	25	1	8	15	22	29	6	13	20	27	3	10	17
Análisis de Requisitos	3 - 60																				
Diseño	3 - 60																				
Implementación	5 - 90																				
Pruebas	5 - 90																				

Tabla 1.1: Planificación inicial del proyecto

La planificación inicial se detalla en la tabla 1.1, sin embargo, experimentó varios retrasos, principalmente debido a que el autor también estaba trabajando en un proyecto en colaboración con la Universidad de Granada. Por otro lado, la obtención de los conjuntos de datos 3D no resultó ser una tarea sencilla, debido a su escasa disponibilidad y a los permisos necesarios para acceder a ellos. Además, el preprocesamiento de los datos 3D consumió

Tarea	Semanas - Horas	Febrero				Marzo				Abril				Mayo				Junio			
		5	12	19	26	4	11	18	25	1	8	15	22	29	6	13	20	27	3	10	17
Análisis de Requisitos	3 - 60																				
Diseño	4 - 70																				
Implementación	6 - 100																				
Pruebas	6 - 100																				

Tabla 1.2:

más tiempo del previsto, ya que, aunque se automatizó en cierta medida, requirió ajustes manuales significativos. Estos contratiempos, junto con el aprendizaje de nuevas librerías por parte del autor, resultaron en modificaciones en la planificación original, tal como se exemplifica en la tabla 1.2.

Capítulo 2

Fundamentos teóricos

2.1. Aprendizaje automático

El aprendizaje automático (*Machine Learning*, ML) [23, 7] es una rama de la inteligencia artificial y de las ciencias de la computación centrada en el uso de datos y algoritmos para imitar la forma en la que los humanos aprenden, detectando patrones o regularidades para realizar predicciones.

Existen 3 tipos de aprendizaje dentro del ML [1, 36]:

El **aprendizaje supervisado** consiste en entrenar un modelo con datos que tienen etiquetas conocidas, lo que indica la categoría a la que pertenece cada dato. Por ejemplo, si los datos de entrada son imágenes de animales, las etiquetas podrían ser 'perro' o 'gato'. A partir de estos datos etiquetados, el modelo aprende a predecir la etiqueta de nuevos datos. Es el tipo de aprendizaje más utilizado y los datos vienen ya 'preparados' para su uso. Es el tipo de aprendizaje que utilizaremos en este TFG.

En el **aprendizaje no supervisado**, el modelo analiza los datos de entrada sin etiquetas, buscando patrones y estructuras inherentes a los datos. El agrupamiento es una técnica común en este tipo de aprendizaje, ya que identifica posibles grupos dentro de los datos. Este enfoque suele requerir un gran volumen de datos para ser efectivo.

Por otro lado, en el **aprendizaje por refuerzo**, el modelo aprende a través de recompensas o penalizaciones en función de las acciones que realiza. El objetivo del agente es maximizar las recompensas a largo plazo, lo que lo hace especialmente útil en la enseñanza de estrategias en juegos y otras interacciones dinámicas.

2.2. Aprendizaje profundo

2.2.1. Redes neuronales artificiales

Las redes neuronales artificiales (*Artificial Neural Networks*, ANN) [17, 20, 5] son redes computacionales que intentan, a groso modo, simular el proceso de decisión de las neuronas del sistema nervioso central de animales y humano. Las ANN poseen unidades de procesamiento de información llamadas neuronas, las cuales están conectadas entre sí. La estructura básica de una ANN se compone de (ver Figura 4.3):

- Una capa de entrada, que tendrá tantos *inputs* como características o variables tenga el problema
- Una o varias capas ocultas, compuestas por neuronas. El número de capas ocultas define la profundidad de la red neuronal.
- Una capa de salida, la cual representa el valor o valores predichos

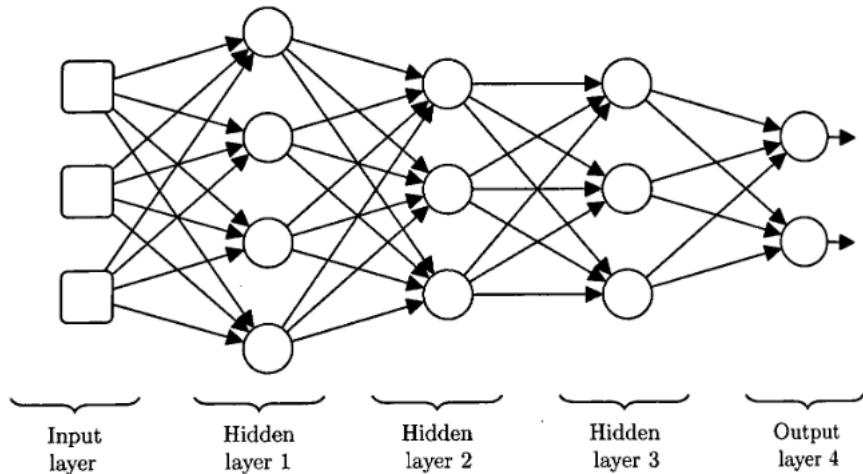


Figura 2.1: Esquema de una red neuronal [24].

Las neuronas son la unidad fundamental de cómputo, tienen varios valores de entrada y un valor de salida que se conecta con las neuronas de la siguiente capa. Los elementos básicos del modelo neuronal son (ver Figura 4.4):

- Un conjunto de conexiones con las señales de entrada. Cada conexión tiene su propio peso/fuerza.
- Una función de suma de las señales de entrada, ponderadas cada una con su peso. Estas operaciones constituyen una combinación lineal.

- Una función de activación, para limitar la amplitud de la salida de la neurona. Normalmente, el rango de salida está en el intervalo [0,1], o alternativamente en [-1,1]. Existen muchos tipos de funciones de activación pero, se suelen utilizar cuatro: la función signo, la función logística, la función arco-tangente o la función ReLU (*Rectified Linear Unit*).

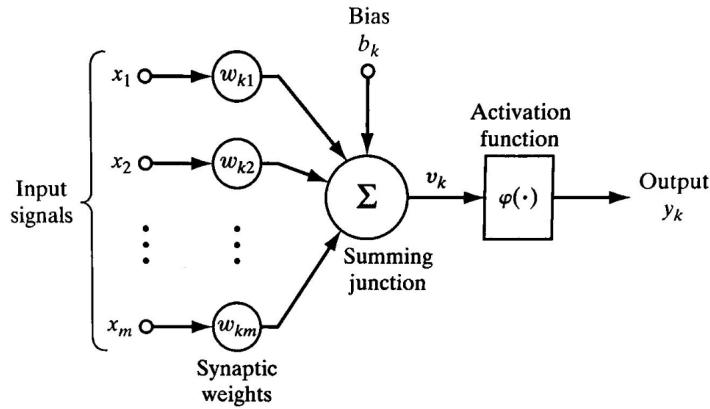


Figura 2.2: Modelo neuronal para una neurona k [20].

En términos matemáticos, podemos describir la salida de una neurona como:

$$y = \phi\left(\sum_{j=1}^m w_j x_j + b\right) \quad (2.1)$$

siendo ϕ la función de activación, m el número de señales de entrada, w_j el peso de cada entrada x_j , y b el sesgo.

El algoritmo de aprendizaje de la red neuronal consiste en ir modificando los pesos y el sesgo, iterativamente, hasta alcanzar el resultado deseado. Este proceso iterativo se conoce como entrenamiento, y permite, a través de las modificaciones de los pesos, reconocer y extraer las características más relevantes de los datos.

El objetivo del entrenamiento es minimizar el error de predicción de la salida de la red neuronal, para ello, se define una función de pérdida. Existen numerosas funciones de pérdida, algunas de las más conocidas son: el error cuadrático medio (MSE), el error absoluto medio (MAE) o la entropía cruzada. La información de la función de pérdida se transmite desde la salida a la capa inicial, con el fin de modificar adecuadamente los pesos para generar una mejor estimación de la predicción.

El sobreentrenamiento es un factor importante a evitar. Sobreentrenar el modelo de aprendizaje, significa, ajustarlo demasiado al conjunto de datos de entrenamiento, de manera que, al recibir nuevos datos no utilizados para entrenar, se estime un mal resultado debido a la poca capacidad de generalización ante nuevos datos.

2.2.2. Redes neuronales convolucionales

Las redes neuronales convolucionales (*Convolutional Neural Networks*, CNN) [26, 27, 49] son un tipo de red neuronal profunda que trabaja con patrones de cuadrícula, como pueden ser imágenes (ver Figura 4.5).

En estas redes neuronales, las capas convolucionales desempeñan un papel fundamental, y a menudo se complementan con capas de *pooling*. Dichas capas se encuentran en la primera parte de la red y son las encargadas de extraer las características relevantes de la entrada. Esto posibilita la automatización del proceso de extracción de características, mejorando simultáneamente tanto el tiempo como el rendimiento.

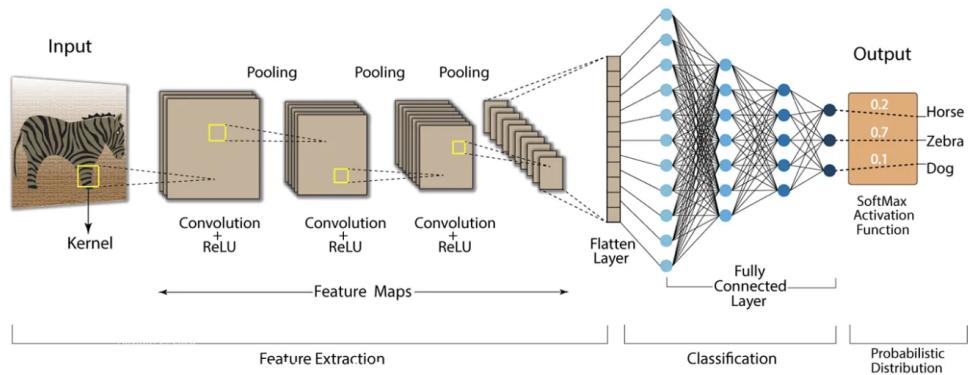


Figura 2.3: Ejemplo de CNN [37].

A continuación, se describen los posibles tipos de capas presentes en una CNN [44, 48]:

Capa de convolución

La capa convolucional es un componente fundamental de las CNN, utilizada para la extracción de características de una imagen o un conjunto de imágenes. Esta capa aplica una operación lineal especializada conocida como convolución, que consiste en aplicar un filtro o *kernel* a la imagen de entrada. El *kernel* es una matriz que se desliza a lo largo de la imagen, multiplicando sus valores con los píxeles correspondientes y sumándolos para

producir un único valor en la imagen de salida. Este proceso se repite en todas las posiciones de la imagen dando como resultado una nueva matriz denominada mapa de características (ver Figura 2.4).

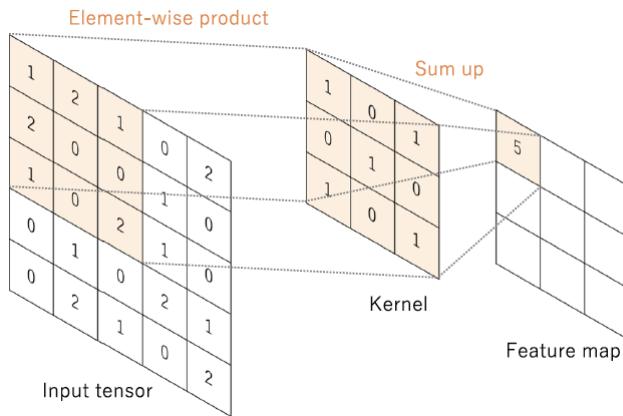


Figura 2.4: Ejemplo de convolución en CNN [48].

Los pesos de los filtros se aprenden durante el proceso de entrenamiento de la red neuronal. Cada *kernel* tiene sus propios pesos que se ajustan iterativamente durante el entrenamiento para minimizar la función de pérdida y mejorar el rendimiento del modelo.

La característica clave de la operación de convolución es el *weight sharing*, que implica compartir los mismos *kernels* en toda la imagen. Esto permite que la red detecte patrones locales independientemente de su ubicación en la imagen. Además, contribuye a aprender jerarquías de características espaciales, lo que permite capturar una amplia gama de características en varios niveles de abstracción. Este enfoque también aumenta la eficiencia del modelo al reducir la cantidad de parámetros que necesita aprender en comparación con las redes totalmente conectadas.

Por otro lado, es importante la configuración de los hiperparámetros de cada capa convolucional, estos se definen antes de iniciar el entrenamiento de la red neuronal y afectan al comportamiento de la misma. Los más comunes son:

- Tamaño del *kernel*: se refiere a las dimensiones del filtro que se aplica a la imagen de entrada. Los tamaños comunes son 3x3, 5x5 o 7x7.
- Número de *kernels*: indica cuántos filtros se aplicarán a la imagen de entrada para extraer diferentes características. Cuantos más *kernels* se utilicen, mayor será la profundidad de los mapas de características de salida.
- *Padding*: esta técnica consiste en añadir píxeles alrededor de la imagen

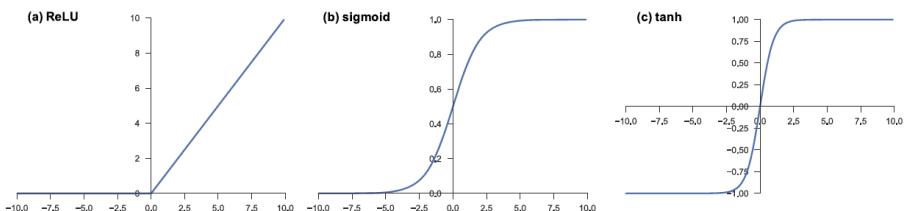
de entrada tras el proceso de convolución. Su propósito es mantener el tamaño de la salida, ya que al aplicar la convolución, las dimensiones del mapa de características se reducen con respecto a la imagen original.

- *Stride*: es el número de píxeles que se desplaza el *kernel* en cada paso durante la convolución. Un mayor *stride* reduce el tamaño del mapa de características y la cantidad de operaciones necesarias.

Sin embargo, es importante tener en cuenta que la operación de convolución por sí sola es lineal y puede no ser suficiente para aprender patrones complejos. En este contexto, entra en juego la capa de activación, que introduce no linealidades en la red y potencia su capacidad para capturar relaciones más complejas entre las características extraídas.

Capa de activación

La capa de activación en una CNN sigue a la capa convolucional y se encarga de introducir no linealidades en el modelo mediante una función de activación. Esta función aumenta la capacidad de la red para aprender relaciones no lineales en los datos, lo que es fundamental para capturar patrones más complejos. Algunas de las funciones de activación comunes utilizadas son la función ReLU, la función sigmoide y la función tangente hiperbólica (ver Figura 2.5).



Capa de *pooling*

La capa de *pooling* también es específica de las CNN y se encarga de reducir la dimensionalidad de las características conservando la información más relevante.

Esta capa resume la información en regiones locales mediante una operación de *downsampling* en las características de entrada. Al reducir la dimensionalidad de las características, la capa de *pooling* disminuye el número

de parámetros aprendibles en la red, lo que puede ayudar a prevenir el sobreajuste y mejorar la eficiencia computacional del modelo. Además, esta capa también ayuda a introducir invariancia a pequeñas traslaciones y distorsiones en los datos de entrada, permitiendo a la red reconocer patrones incluso si están ligeramente desplazados en la imagen.

Los dos tipos más comunes de *pooling* son el *max pooling*, que selecciona el valor máximo de una región local en las características de entrada, y el *average pooling*, que calcula el promedio de los valores en una región local (ver Figura 2.6).

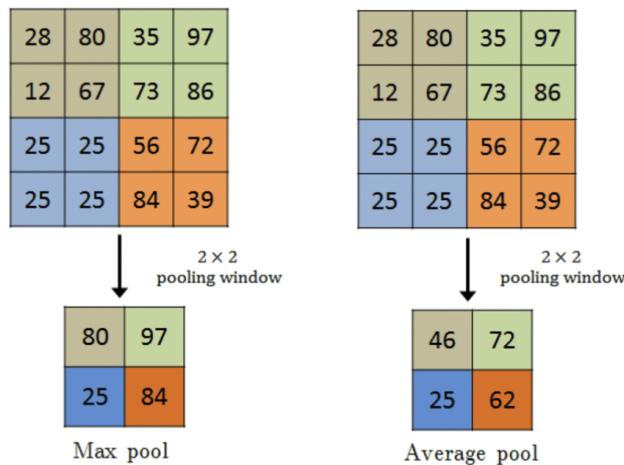


Figura 2.6: Tipos de *pooling* comúnmente utilizados en CNN [48].

Los hiperparámetros de la capa de *pooling* incluyen el tamaño del filtro, el *stride* y el tipo de *padding*. Estos hiperparámetros afectan la forma en que se realiza el *downsampling* en las características.

Capa totalmente conectada

La capa totalmente conectada sigue a las capas de convolución y *pooling*. En esta capa, las características extraídas por las capas anteriores se transforman en un formato unidimensional (vector) antes de conectarse a una o más capas totalmente conectadas, también conocidas como capas densas.

Esta capa, al igual que en las redes neuronales clásicas, tiene la responsabilidad de combinar y procesar las características extraídas para producir la salida final de la red.

Cada neurona en una capa totalmente conectada está conectada a todas las neuronas de la capa anterior a través de pesos aprendibles. Estos pesos determinan la contribución de cada neurona de entrada a la neurona de salida correspondiente en la capa totalmente conectada. Durante el entre-

namiento, estos pesos se ajustan mediante algoritmos de optimización como *backpropagation* y descenso de gradiente para minimizar la diferencia entre las salidas predichas y las etiquetas reales.

Es importante destacar que la capa totalmente conectada suele estar seguida por una función de activación no lineal, como ReLU, para introducir no linealidades en el modelo y permitir la representación de patrones complejos en los datos. Además, la última capa de activación de la CNN, generalmente se selecciona según la naturaleza de la tarea que se está abordando.

2.2.3. Transferencia de aprendizaje

La transferencia de aprendizaje [21], también conocida como *Transfer Learning* (TL), es una técnica fundamental en el campo del aprendizaje automático. Consiste en aprovechar el conocimiento adquirido al resolver un problema para mejorar el rendimiento en otro problema relacionado. En lugar de comenzar desde cero al entrenar un modelo para una tarea específica, el TL utiliza el aprendizaje previo en tareas similares, obteniendo múltiples beneficios, como una mayor eficiencia en el entrenamiento de modelos, una mejor generalización con conjuntos de datos limitados y una aceleración en el desarrollo de modelos.

En las arquitecturas convolucionales, la forma más común de llevar a cabo la transferencia de aprendizaje es mediante el *fine-tuning* [44], que implica utilizar pesos pre-entrenados, congelar todas las capas de la red excepto las superiores, y ajustar estas últimas para adaptarlas a nuestro problema específico, de manera que el entrenamiento se realice únicamente en esas capas superiores. Este enfoque aprovecha la capacidad de los modelos pre-entrenados para capturar características generales de los datos, lo cual es especialmente útil cuando se dispone de conjuntos de datos pequeños o limitados. Además, al congelar las capas iniciales se evita la pérdida de información importante aprendida durante el pre-entrenamiento, mientras que el *fine-tuning* en las capas superiores permite adaptar el modelo a la nueva tarea específica.

En este contexto, es común utilizar los pesos pre-entrenados en el conjunto de datos de ImageNet [22] debido a su gran tamaño, diversidad, representatividad y disponibilidad.

2.2.4. Regularización

Tanto en las redes neuronales clásicas como en las convolucionales, el sobreajuste a los datos de entrenamiento es una problema importante (ver Figura 2.7). Aunque la solución óptima sería adquirir más datos para el entrenamiento, esta opción no siempre está disponible. Por tanto, se recu-

rre a técnicas de regularización para mitigar este problema. Entre las más destacadas se encuentran:

- *Dropout*: es una técnica de regularización donde se establecen aleatoriamente ciertas activaciones a 0 durante el entrenamiento, de modo que el modelo se vuelve menos sensible a pesos específicos en la red.
- *Weight decay*: también conocido como regularización L2, reduce el sobreajuste penalizando los pesos del modelo para que tomen solo valores pequeños.
- *Batch normalization*: es un tipo de capa suplementaria que normaliza adaptativamente los valores de entrada de la siguiente capa, mitigando el riesgo de sobreajuste, así como mejorando el flujo de gradiente a través de la red, permitiendo tasas de aprendizaje más altas y reduciendo la dependencia de la inicialización.
- *Data augmentation*: es un proceso de modificación de los datos de entrenamiento a través de transformaciones aleatorias, como volteo, traslación, recorte, rotación y borrado aleatorio, para que el modelo no vea exactamente las mismas entradas durante las iteraciones de entrenamiento. Esta técnica, además de reducir el sobreajuste, permite una mejor generalización del modelo.
- Elección del modelo: un modelo de una alta complejidad puede provocar sobreajuste ya que tiene la capacidad de ajustarse mucho mejor a los datos de entrenamiento. Es fundamental encontrar un modelo que tenga un equilibrio entre complejidad y generalización, es decir, que sea lo suficientemente complejo para captar las características importantes pero que a la vez sea capaz de generalizar sin sobreajustarse demasiado a los datos.

A pesar de las técnicas anteriores, persiste la preocupación por el sobreajuste al conjunto de validación en lugar del conjunto de entrenamiento, principalmente debido a la filtración de información durante el ajuste fino de hiperparámetros y el proceso de selección del modelo. Por tanto, es importante evaluar el rendimiento del modelo final en un conjunto de prueba separado, preferiblemente no visto previamente. Esto es fundamental para validar la capacidad de generalización del modelo y garantizar su fiabilidad.

2.3. Parámetros de la cámara y perspectiva

Dado que el conjunto de datos utilizado en este TFG son imágenes creadas sintéticamente, es importante conocer los parámetros de la cámara que influyen a la hora de sacar las fotografías.

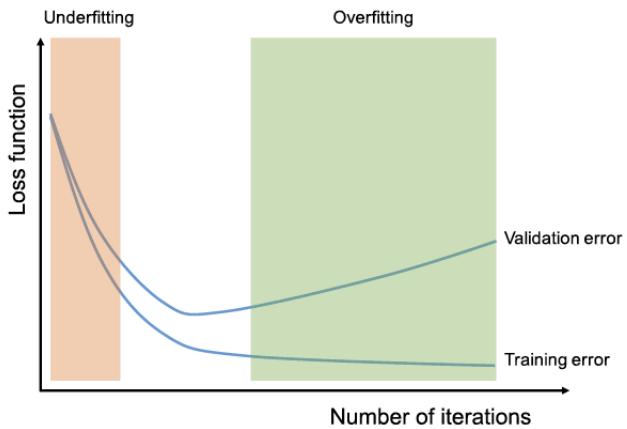


Figura 2.7: Zona de infraajuste y sobreajuste durante el entrenamiento [48].

Longitud focal

La longitud focal [18] mide la distancia, en milímetros, entre el 'punto nodal' (punto donde la luz converge en una lente) y el sensor de la cámara (ver Figura 2.8).

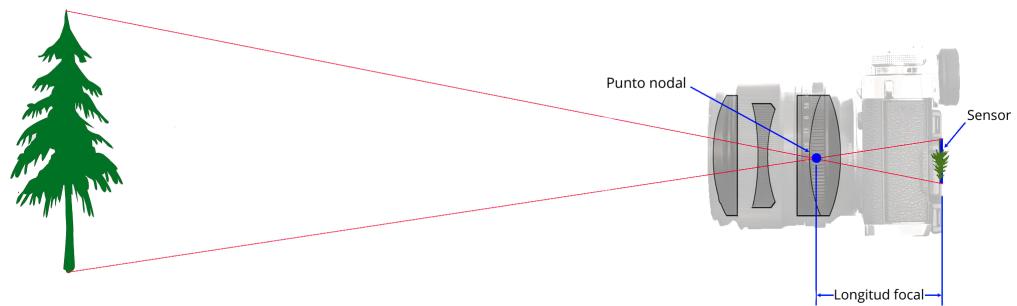


Figura 2.8: Relación entre punto nodal y longitud focal [18]

La longitud focal es importante porque se relaciona con el campo de visión de una lente, es decir, cuánta escena se captura. También explica qué tan grande o pequeño aparecerá un sujeto en la fotografía.

Valores más altos de longitud focal (como 500 mm) se ven más 'cerca', mientras que valores más bajos (como 20 mm) están más 'lejos'.

Sensor de la cámara

El sensor de la cámara [2, 29] es el componente que captura la luz y la convierte en una imagen digital. El tamaño del sensor afecta a la calidad de

la imagen y a la cantidad de luz que puede capturar. El tamaño estándar es de 36mm x 24mm, también conocido como *full frame* o 35mm.

Otra forma equivalente de referirnos al tamaño del sensor es mediante el factor de recorte, que se define como la relación existente entre el tamaño de un sensor de 35mm y el sensor de nuestra cámara (ver Figura 2.9).



Figura 2.9: Tamaños del sensor expresados según el factor de recorte [6]

Uno de los aspectos más importantes del factor de recorte es su influencia en la longitud focal, lo que nos lleva a hablar de 'longitud focal equivalente'. Por ejemplo, al tener una focal de 300 mm en un sensor con factor de recorte 1.6, estaríamos obteniendo un efecto equivalente al de una focal de 480 mm (300 mm x 1.6) en un sensor *full frame* (factor de recorte 1).

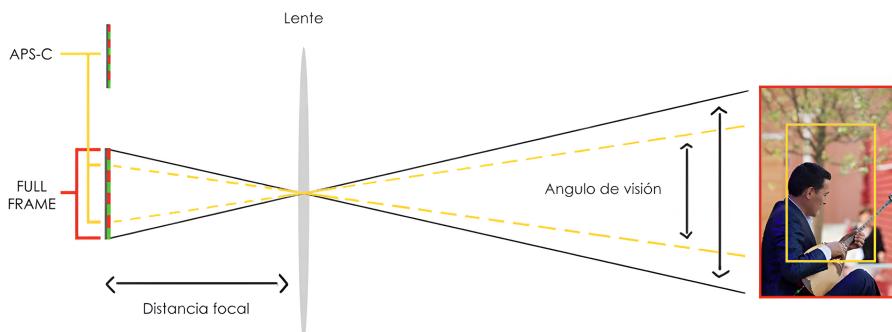


Figura 2.10: Ejemplo de longitud focal equivalente según el tamaño del sensor [41].

Distancia cámara-sujeto

La distancia cámara-sujeto se define como la separación física entre la cámara y el sujeto que está siendo fotografiado (ver Figura 2.11). Modificar esta distancia provoca variaciones en la apariencia visual del rostro en la fotografía obtenida [31]. Este fenómeno se conoce como distorsión de perspectiva.

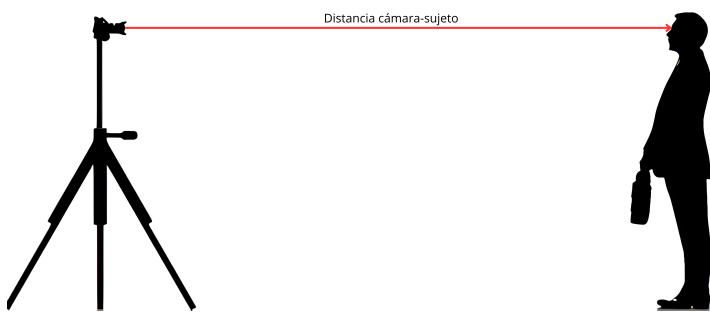


Figura 2.11: Distancia desde la cámara al sujeto.

Distorsión de perspectiva

La distorsión de perspectiva [43, 13] es la transformación que sufre un objeto y su entorno debido a la proximidad del mismo respecto al objetivo (ver Figura 2.12). En el caso de las fotografías faciales, cuanto menor es la distancia cámara-sujeto, mayor es la distorsión de perspectiva que afecta a la persona fotografiada. Esto afecta a rasgos de la cara que pueden aparecer más grandes, como la nariz, o más pequeños, como las orejas, de lo que realmente son.

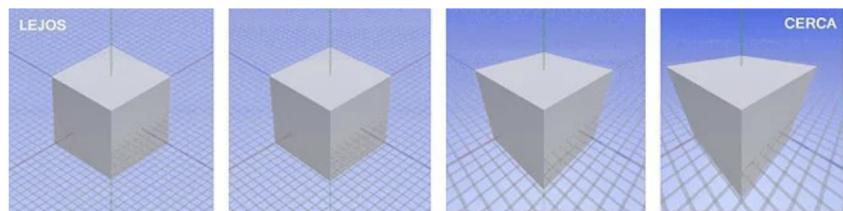


Figura 2.12: Efecto de la distorsión conforme se acerca la cámara al objeto [13].

Uno de los malentendidos comunes en fotografía es la creencia de que la longitud focal distorsiona los rasgos faciales, sin embargo, la longitud focal

no tiene nada que ver con la distorsión del rostro de un sujeto, siendo esta únicamente provocada por la distancia de la cámara al sujeto [30].

Capítulo 3

Estado del Arte

En el campo del aprendizaje automático, el tema de la estimación de la distancia en fotografías faciales ha ganado recientemente mucha atención. Se puede observar en la Figura 4.5 la cantidad de publicaciones existentes en la base de datos Scopus¹ que hacen referencia a la estimación de la SCD. Hay 441 publicaciones registradas desde 1992.

El número de publicaciones relacionadas con este tema, va aumentando a lo largo del tiempo, llegando a obtener un mayor número de publicaciones en 2020. Pese al aumento de publicaciones en este ámbito, es a partir del 2015 cuando se empiezan a aplicar las técnicas de aprendizaje profundo. Este aumento está relacionado con los avances tecnológicos que permiten aplicar nuevas técnicas y conocimientos.

3.1. Primeros enfoques

El primer método utilizado para abordar la estimación métrica de la SCD fue propuesto por Flores et al. [16], quienes proponen utilizar un conjunto de puntos de referencia faciales para calcular la distancia y la posición respecto a la cámara, en un rango que va desde los 10 cm hasta los 3 m. Este método consiste en tomar una imagen 2D de una cara desconocida, identificar sus puntos de referencia faciales y compararlos con los puntos obtenidos de modelos faciales 3D conocidos. Luego, empleando el algoritmo EPnP [28], se determina la distancia entre la cámara y el sujeto. Esta técnica asume que los puntos de referencia no varían significativamente entre individuos, sino que tienden a agruparse en *clusters*.

Sin embargo, este primer enfoque presenta algunas limitaciones, como la dependencia de conjuntos de datos en 3D (los cuales no siempre están

¹Las búsquedas se pueden consultar en el apéndice

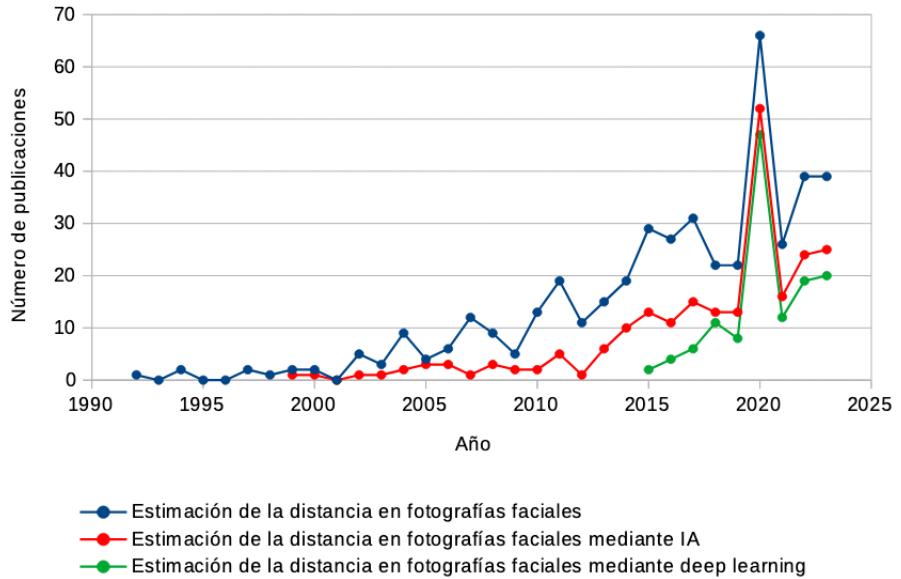


Figura 3.1: Número de publicaciones, en Scopus, relacionadas con la estimación de la distancia en fotografías faciales en función del año de publicación

disponibles), la mezcla de diferentes longitudes focales en un mismo conjunto de datos y la necesidad de reconocimiento manual de los puntos de referencia faciales.

Posteriormente, Burgos-Artizzu et al. [10] introducen un método innovador que elimina la necesidad de anotación manual de los puntos de referencia de la imagen. En su lugar, estos puntos se estiman automáticamente mediante un enfoque de regresión conocido como *Robust Cascaded Pose Regression* (RCPR) [8]. Una vez que se han identificado los puntos de referencia faciales, se emplea un modelo automático de regresión para predecir la distancia entre la cámara y el sujeto en función de la posición relativa de estos puntos. Este regresor fue entrenado utilizando el conjunto de datos Caltech Multi-Distance Portraits (CMDP) [9], que consta de 53 retratos individuales tomados desde 7 distancias diferentes, que van desde 60 cm hasta 480 cm. Todos estos retratos están anotados manualmente con 55 marcas faciales.

Este método, sigue teniendo algunas limitaciones como el recorte de las imágenes (pérdida de resolución) o la única vista frontal.

Además de los métodos previamente citados, se han desarrollado otras técnicas para estimar la SCD basadas en características anatómicas como el tamaño facial [38], la separación entre los ojos [34], o una combinación de ambos factores [25].

3.2. MediaPipe Iris

MediaPipe Iris² es un modelo de aprendizaje automático desarrollado por investigadores de Google. Este modelo tiene la capacidad de rastrear puntos de referencia como el iris, la pupila y los contornos del ojo en tiempo real, utilizando únicamente una cámara RGB estándar y sin necesidad de utilizar ningún hardware especializado. Mediante el seguimiento de los puntos de referencia del iris, este modelo puede determinar la distancia métrica entre el sujeto y la cámara.

El modelo se basa en el diámetro horizontal del iris del ojo humano, el cual se mantiene relativamente constante en un rango de 11.7 ± 0.5 mm en una amplia población. Esta característica, combinada con argumentos geométricos simples, permite al modelo estimar la distancia SCD.

Sin embargo, es importante destacar que este modelo presenta ciertas condiciones y limitaciones. Es útil únicamente en situaciones donde existan datos EXIF disponibles, se capturen imágenes frontales donde el iris sea visible, y los individuos se encuentren a una distancia de menos de 2 metros de la posición de la cámara.

3.3. PerspectiveX

PerspectiveX, desarrollado por Stephan et al. [40], es un método para la estimación de la SCD en imágenes faciales, diseñado para mejorar el proceso de superposición craneofacial.

Este método se basa en la localización de una característica anatómica específica, la longitud de la fisura palpebral, definida por dos puntos de referencia fácilmente identificables.

Esta elección se justifica por varios aspectos: su clara visibilidad frontal, incluso cuando la cabeza experimenta un ligero giro hacia el lado más cercano a la cámara; su definición precisa, que garantiza una correcta medición; su mínima variabilidad, atribuible a restricciones evolutivas; su notable tamaño facial relativo, lo cual minimiza la probabilidad de errores en comparación con características más pequeñas, como el diámetro del iris; y su distribución normal, que contribuye a reducir el margen de error en las predicciones.

Además de la fisura palpebral, PerspectiveX requiere conocer el tipo de cámara, necesario para obtener las especificaciones de píxeles, así como la longitud focal de las lentes. Ambos datos pueden extraerse de las imágenes electrónicas mediante lectores EXIF disponibles en línea.

Finalmente, la estimación del SCD se realiza mediante la siguiente fórmu-

²<https://blog.research.google/2020/08/mediapipe-iris-real-time-iris-tracking.html>

la:

$$SCD = f \left(1 + \frac{A}{x \cdot y} \right) \quad (3.1)$$

donde: f , es la longitud focal de las lentes (mm); A , es la longitud real de la fisura palpebral (mm); x , es la longitud de la fisura palpebral en la foto (píxeles); y , son las especificaciones del tamaño del píxel del receptor de imagen (mm)

Dado que la longitud real de la fisura palpebral puede no estar disponible, se recurre al promedio de un grupo demográfico homogéneo en términos de sexo y edad, ya que se sabe que esta medida varía mínimamente debido a restricciones evolutivas.

Aunque PerspectiveX ofrece una estimación precisa de la SCD para una longitud focal conocida, presenta ciertas limitaciones, como la necesidad de intervención manual para marcar los puntos de referencia faciales y la falta de consideración de las rotaciones de cabeza superiores a 30° .

3.4. FacialSCDnet

FacialSCDnet, presentado por Bermejo et al. [4], es un método que estima la SCD directamente a partir de fotografías mediante el empleo de técnicas de aprendizaje profundo. La utilización de una arquitectura de redes neuronales profundas elimina una restricción crucial: la necesidad de detectar una característica anatómica específica para guiar el proceso de estimación. Esta capacidad permite que el método sea eficaz en la estimación de la SCD en cualquier posición de la cabeza, desde la frontal hasta el perfil lateral.

Para entrenar el modelo, se empleó un conjunto de datos compuesto por dos colecciones:

- Conjunto sintético: se generaron imágenes sintéticas 2D a partir de los modelos 3D de la base de datos Stirling ESRC 3D Face³. En particular, se utilizaron 315 modelos faciales de 54 individuos diferentes para generar aproximadamente 150.000 fotografías sintéticas.
- Conjunto de fotografías digitales: se adquirieron fotografías de 28 individuos siguiendo un protocolo de adquisición específico. Se consideraron 4 longitudes focales diferentes (27 mm, 35 mm, 55 mm, 85 mm) en formato full frame y se capturaron 12 distancias diferentes de la

³Stirling ESRC 3D Face: <https://pics.stir.ac.uk/ESRC/index.htm>

cámara al sujeto, que oscilaron desde 50 cm hasta 6 m. Además, se fotografiaron 7 posiciones distintas de la cabeza, desde el perfil izquierdo hasta el perfil derecho, con intervalos de rotación de 30º.

La función de pérdida empleada en el modelo se basa en el error absoluto medio de la distorsión facial relativa, calculada mediante la fórmula:

$$Distortion = \frac{\sum_{i=1}^n |y_i - x_i|}{n} \quad (3.2)$$

donde y_i son los valores reales (etiquetas) de la distorsión facial, y x_i son los valores predichos de la distorsión facial, calculados a partir del factor de distorsión (D_f):

$$D_f = \frac{1}{1 + \frac{SCD}{d}} \quad (3.3)$$

En la ecuación 3.3, $d = 12.6572$ cm corresponde a un valor derivado de cálculos geométricos [40] para obtener experimentalmente el factor de distorsión de una cabeza humana de tamaño promedio, según la SCD de la fotografía.

FacialSCDnet consta de 4 modelos de aprendizaje profundo, cada uno asociado a una longitud focal utilizada en el conjunto de datos. La estructura de cada CNN se basa en la arquitectura VGG-16, cuyos pesos se inicializan con los pre-entrenados en ImageNet ⁴. Para adaptar la arquitectura al problema de estimación de la SCD, se conservan los 5 bloques convolucionales, se elimina la capa superior y se añaden 2 capas totalmente conectadas que se entrena desde cero. Finalmente, la última capa del modelo consiste en una activación lineal que realiza la tarea de regresión.

El proceso de entrenamiento de los modelos constó de dos fases. En primer lugar, los modelos se entrenaron con el conjunto de datos sintético para aprender las relaciones entre la SCD y las características faciales. Posteriormente, se realizó un ajuste fino utilizando el conjunto de datos reales.

Los resultados obtenidos indican que las cuatro redes de FacialSCDnet son capaces de predecir con precisión la SCD, con errores promedio por debajo de 5 cm (MAE) o 3% (MRE). Esta precisión en la predicción de la distancia métrica se traduce en un error promedio del 0.2% al considerar la métrica de distorsión facial relativa. Estos resultados muestran que FacialSCDnet logra una estimación precisa de la SCD en fotografías faciales, superando a otros métodos existentes y demostrando su robustez y eficacia en diversas situaciones y condiciones.

⁴ImageNet: <https://www.image-net.org/>

Capítulo 4

Materiales y métodos

4.1. Materiales

4.1.1. Modelos 3D

Para este TFG, se llevó a cabo un análisis exhaustivo de las bases de datos disponibles de modelos 3D de personas. Tras la búsqueda, se utilizaron diversos conjuntos de datos públicos con el objetivo de crear un conjunto de datos unificado, realista y diverso. En este conjunto, se incluyeron tanto modelos faciales como modelos de cuerpo completo.

Modelos faciales

Se han seleccionado los siguientes conjuntos de datos: HeadSpace [12], H3DS-net [35] y DI4D_UGR_ANON ¹.

El conjunto de datos de Headspace [12] es un conjunto de imágenes en 3D de la cabeza humana, que consta de 1519 sujetos que llevan gorros de látex ajustados para reducir el efecto de los peinados. Las ventajas de este conjunto son que tienen muy buena resolución y además incluye metadatos útiles para seleccionar un subconjunto de datos adecuado.

H3DS-net [35] contiene escaneos texturizados en 3D de la cabeza completa con una alta resolución. Este conjunto comprende un total de 23 modelos, todos ellos con los ojos cerrados, lo que añade una variabilidad adicional.

DI4D_UGR_ANON es un conjunto de datos creado por la Universidad de Granada...

Si bien se investigaron otros conjuntos de datos como FaceVerse [45] o

¹Conjunto de datos proporcionado por el tutor

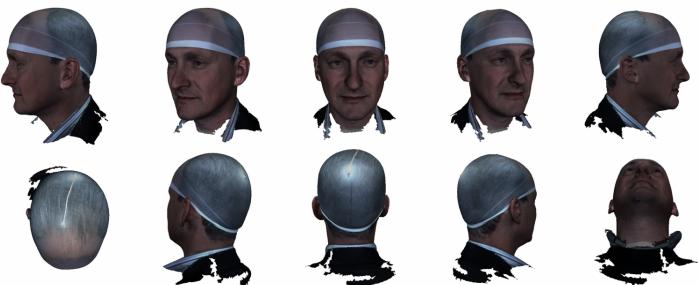


Figura 4.1: HeadSpace 3D



Figura 4.2: H3DS-net

CASIA², estos fueron descartados debido a problemas como la baja calidad, formatos incompatibles y escalas no reales.

Modelos de cuerpo entero

Se han seleccionado los siguientes conjuntos de datos: HuMMan [11], People Snapshot [3] y Render People³



Figura 4.3: HuMMan

²<http://biometrics.idealtest.org/>

³<https://renderpeople.com/es/>



Figura 4.4: People Snapshot



Figura 4.5: Render People

Selección del subconjunto de datos

4.1.2. Procesamiento del conjunto de datos

En este apartado explicaremos la necesidad de alinear los modelos 3D para tener el origen a la altura de los ojos y así poder tener una referencia a la hora de generar imágenes 2D a distintas distancias. Se expondrán los métodos llevados a cabo para dicha finalidad (tanto alinear como generar las imágenes). Además, se explicará cómo se han cambiado los fondos e iluminación de las imágenes para hacerlas más realistas.

4.2. Métodos

Alineamiento de modelos 3D

Generación de fotografías faciales a partir de modelos 3D

Mejoras en fondo e iluminación de imágenes

En los siguientes apartados se describen las arquitecturas de deep learning que vamos a utilizar para realizar los experimentos.

4.2.1. FacialSCDnet+

Bibliografía

- [1] Yaser S. Abu-Mostafa, M. Magdon-Ismail y H.T. Lin. *Learning from Data: A Short Course*. AMLBook, 2012.
- [2] Adorama. *What Is Crop Factor And How Do You Calculate It?* Accedido el 17 de Marzo de 2024. 2022. URL: <https://www.adorama.com/alc/what-is-crop-factor-everything-you-need-to-know/>.
- [3] Thiemo Alldieck et al. «Video Based Reconstruction of 3D People Models». En: *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2018, págs. 8387-8397.
- [4] Enrique Bermejo et al. «FacialSCDnet: A deep learning approach for the estimation of subject-to-camera distance in facial photographs». En: *Expert Systems with Applications* 210 (2022), pág. 118457.
- [5] C.M. Bishop. *Neural networks for pattern recognition*. Oxford University Press, 1995.
- [6] Blas. *Factor de recorte de un sensor*. Accedido el 18 de Marzo de 2024. 2018. URL: <https://blasfotografia.com/factor-de-recorte-de-un-sensor/>.
- [7] Sara Brown. *Machine learning, explained*. Accedido el 11 de Marzo de 2024. 2021. URL: <https://mitsloan.mit.edu/ideas-made-to-matter/machine-learning-explained>.
- [8] Xavier Burgos-Artizzu, Pietro Perona y Piotr Dollar. «Robust Face Landmark Estimation under Occlusion». En: *IEEE International Conference on Computer Vision* (2013), págs. 1513-1520.
- [9] Xavier Burgos-Artizzu, Matteo Ruggero Ronchi y Pietro Perona. *Caltech Multi-Distance Portraits (CMDP)*. 2022.
- [10] Xavier P. Burgos-Artizzu, Matteo Ruggero Ronchi y Pietro Perona. «Distance Estimation of an Unknown Person from a Portrait». En: *European Conference on Computer Vision*. Vol. 8689. 2014, págs. 313-327.
- [11] Zhongang Cai et al. «HuMMan: Multi-modal 4d human dataset for versatile sensing and modeling». En: *European Conference on Computer Vision*. 2022, págs. 557-577.

- [12] Hang Dai et al. «Statistical Modeling of Craniofacial Shape and Texture». En: *International Journal of Computer Vision* 128.2 (2019), págs. 547-571.
- [13] Alfonso Domínguez. *Distorsión de lente vs Distorsión de la perspectiva*. Accedido el 23 de Marzo de 2024. 2011. URL: <https://www.xatakafoto.com/guias/distorsion-de-lente-vs-distorsion-de-la-perspectiva>.
- [14] Gary Edmond et al. «Law's Looking Glass: Expert Identification Evidence Derived from Photographic and Video Images». En: *Current Issues in Criminal Justice* 20 (2009), págs. 337-377.
- [15] FISWG. *Facial Comparison Overview and Methodology Guidelines V2.0*. Accedido el 7 de Febrero de 2024. 2022. URL: https://fiswg.org/fiswg_facial_comparison_overview_and_methodology_guidelines_V2.0_2022.11.04.pdf.
- [16] Arturo Flores et al. «Camera Distance from Face Images». En: *Advances in Visual Computing*. Vol. 8034. 2013, págs. 513-522.
- [17] Daniel Graupe. *Principles of artificial neural networks 3rd edition*. World Scientific, 2007.
- [18] Elizabeth Gray. *What Is Focal Length in Photography? A Beginner's Guide*. Accedido el 17 de Marzo de 2024. 2023. URL: <https://photographylife.com/what-is-focal-length-in-photography>.
- [19] Ankush Gupta, Andrea Vedaldi y Andrew Zisserman. «Synthetic Data for Text Localisation in Natural Images». En: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* 2016 (2016), págs. 2315-2324.
- [20] Simon Haykin. *Neural Networks and Learning Machines 3rd edition*. Pearson, 2009.
- [21] Asmaul Hosna et al. «Transfer learning: a friendly introduction». En: *Journal of Big Data* 9 (2022).
- [22] Minyoung Huh, Pulkit Agrawal y Alexei Efros. «What makes ImageNet good for transfer learning?» En: (2016).
- [23] IBM. *What is machine learning?* Accedido el 11 de Marzo de 2024. 2019. URL: <https://www.ibm.com/topics/machine-learning>.
- [24] John D. Kelleher. *Deep learning*. MIT Press, 2019.
- [25] M.S. Shashi Kumar, K.S. Vimala y N. Avinash. «Face distance estimation from a monocular camera». En: *IEEE International Conference on Image Processing*. 2013, págs. 3532-3536.
- [26] Y. LeCun et al. «Backpropagation Applied to Handwritten Zip Code Recognition». En: *Neural Computation* 1 (1989), págs. 541-551.

- [27] Y. Lecun et al. «Gradient-based learning applied to document recognition». En: *Proceedings of the IEEE* 86 (1998), págs. 2278-2324.
- [28] Vincent Lepetit, Francesc Moreno-Noguer y Pascal Fua. «EPnP: An accurate O(n) solution to the PnP problem». En: *International Journal of Computer Vision* 81 (2009).
- [29] Javier Lucas. *Qué Es El Factor de Recorte de tu Sensor y Cómo Influye en la Focal de tus Objetivos*. Accedido el 17 de Marzo de 2024. 2022. URL: <https://www.dzoom.org.es/que-es-el-factor-de-recorte-de-tu-sensor-y-como-influye-en-la-focal-de-tus-objetivos/>.
- [30] Nasim Mansurov. *Does Focal Length Distort Subjects?* Accedido el 23 de Marzo de 2024. 2020. URL: <https://photographylife.com/does-focal-length-distort-subjects>.
- [31] Eilidh Noyes y Rob Jenkins. «Camera-to-subject distance affects face configuration and perceived identity». En: *Cognition* 165 (2017), págs. 97-104.
- [32] Bo Peng et al. «Position Determines Perspective: Investigating Perspective Distortion for Image Forensics of Faces». En: *IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 2017, págs. 1813-1821.
- [33] R.S. Pressman. *Software Engineering: A Practitioner's Approach*. McGraw-Hill, 2005.
- [34] Khandaker Abir Rahman et al. «Person to Camera Distance Measurement Based on Eye-Distance». En: *International Conference on Multimedia and Ubiquitous Engineering*. 2009, págs. 137-141.
- [35] Eduard Ramon et al. «H3D-Net: Few-Shot High-Fidelity 3D Head Reconstruction». En: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, págs. 5620-5629.
- [36] Stuart Russell y Peter Norvig. *Artificial Intelligence: A Modern Approach 4^ºedition*. Pearson, 2021.
- [37] Nafiz Shahriar. *What is Convolutional Neural Network — CNN (Deep Learning)*. Accedido el 12 de Marzo de 2024. 2023. URL: <https://nafizshahriar.medium.com/what-is-convolutional-neural-network-cnn-deep-learning-b3921bdd82d5>.
- [38] Mohamed Tahir Ahmed Shoani, Shamsudin H. M. Amin e Ibrahim M. H. Sanhoury. «Determining subject distance based on face size». En: *Asian Control Conference*. 2015, págs. 1-6.
- [39] Nicole Spaun. «Facial Comparisons by Subject Matter Experts: Their Role in Biometrics and Their Training». En: *International Conference on Biometrics*. 2009, págs. 161-168.

- [40] Carl N. Stephan. «Estimating the Skull-to-Camera Distance from Facial Photographs for Craniofacial Superimposition». En: *Journal of Forensic Sciences* 62 (2017), págs. 850-860.
- [41] Nicholas Tinelli. *Sensores y factor de recorte: en palabras fáciles*. Accedido el 17 de Marzo de 2024. 2020. URL: <https://nicholastinelli.com/es/sensores-y-factor-de-recorte-en-palabras-faciles/>.
- [42] Jonathan Tremblay et al. «Training deep networks with synthetic data: Bridging the reality gap by domain randomization». En: *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops* 2018 (2018), págs. 1082-1090.
- [43] Andrea Valsecchi. *Comprendión de los parámetros de la cámara*. Accedido el 17 de Febrero de 2024. 2019. URL: <https://skeleton-id.com/investigaciones/comprehension-de-los-parametros-de-la-camara>.
- [44] Gal Vardi. «On the Implicit Bias in Deep-Learning Algorithms». En: *Communications of the Association for Computing Machinery* 66 (2022), págs. 86-93.
- [45] Lizhen Wang et al. «FaceVerse: a Fine-grained and Detail-controllable 3D Face Morphable Model from a Hybrid Dataset». En: *IEEE Conference on Computer Vision and Pattern Recognition*. 2022.
- [46] Qi Wang et al. «Learning from synthetic data for crowd counting in the wild». En: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* 2019 (2019), págs. 8190-8199.
- [47] Zhixiang Wang et al. «DisCO: Portrait Distortion Correction with Perspective-Aware 3D GANs». En: (2023).
- [48] Rikiya Yamashita et al. «Convolutional neural networks: an overview and application in radiology». En: *Insights into Imaging* 9 (2018), págs. 611-629.
- [49] Guangle Yao, Tao Lei y Jiandan Zhong. «A review of Convolutional-Neural-Network-based action recognition». En: *Pattern Recognition Letters* 118 (2019), págs. 14-22.
- [50] Xiangxin Zhu y Deva Ramanan. «Face Detection, Pose Estimation, and Landmark Localization in the Wild». En: *IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (2012), págs. 2879-2886.

Búsquedas en Scopus

Imágenes faciales (2937)

TITLE-ABS-KEY (facial AND (images OR photographs))

Estimación de la distancia en fotografías faciales (441)

TITLE-ABS-KEY ((distance OR depth) AND estimation AND (photographs OR images) AND facial) AND (LIMIT-TO (SUBJAREA , ÇOMP") OR LIMIT-TO (SUBJAREA , .^{ENGI}"))

Estimación de la distancia en fotografías faciales mediante IA (224)

TITLE-ABS-KEY ((deep AND learning) OR (machine AND learning) OR (artificial AND intelligence) OR (computer AND vision) OR (soft AND computing) AND ((distance OR depth) AND estimation AND (photographs OR images) AND facial)) AND (LIMIT-TO (SUBJAREA , ÇOMP") OR LIMIT-TO (SUBJAREA , .^{ENGI}"))

Estimación de la distancia en fotografías faciales mediante deep learning (129)

TITLE-ABS-KEY ((deep AND learning) AND ((distance OR depth) AND estimation AND (photographs OR images) AND facial)) AND (LIMIT-TO (SUBJAREA , ÇOMP") OR LIMIT-TO (SUBJAREA , .^{ENGI}"))