TRABAJO FIN DE GRADO

INGENIERÍA INFORMÁTICA

# Estimación de distancia cámara-sujeto en fotografías faciales usando deep learning


**Autor**
Iván Salinas López

**Directores**
Pablo Mesejo Santiago
Enrique Bermejo Nievas



Escuela Técnica Superior de Ingenierías Informática y de Telecomunicación


—

Granada, Julio de 2023

**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 TITULACIÓN 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 Julio de 2023

D. **Nombre Apellido1 Apellido2 (tutor1)**, Profesor del Área de XXXX del Departamento YYYY de la Universidad de Granada.

D. **Nombre Apellido1 Apellido2 (tutor2)**, Profesor del Área de XXXX del Departamento YYYY de la Universidad de Granada.

**Informan:**

Que el presente trabajo, titulado ***Título del proyecto, Subtítulo del proyecto***, ha sido realizado bajo su supervisión por **Nombre Apellido1 Apellido2 (alumno)**, 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 mes de 201 .

**Los directores:**

**Nombre Apellido1 Apellido2 (tutor1)** **Nombre Apellido1 Apellido2 (tutor2)**

# Agradecimientos

Poner aquí agradecimientos...

# Índice general

**1. Introducción** **1**
1.1. Definición del problema . . . . . 1
1.2. Motivación . . . . . 3
1.3. Objetivos . . . . . 4

**2. Fundamentos teóricos** **5**
2.1. Aprendizaje automático . . . . . 5
2.2. Aprendizaje profundo . . . . . 6
2.2.1. Redes neuronales . . . . . 6
2.2.2. Redes neuronales convolucionales . . . . . 8

**3. Estado del Arte** **9**
3.1. Estimación de distancia cámara-sujeto . . . . . 9
3.2. FacialSCDnet . . . . . 9

**4. Materiales y métodos** **11**
4.1. Materiales . . . . . 11
4.2. Métodos . . . . . 11

**5. Experimentos** **13**
5.1. ... . . . . 13
5.2. ... . . . . 13
5.3. ... . . . . 13

**6. Conclusiones y trabajos futuros** **15**
6.1. ... . . . . 15
6.2. ... . . . . 15
6.3. ... . . . . 15

# Índice de figuras

1.1. Efectos de la distorsión de perspectiva en características faciales de fotografías realizadas a diferentes SCD: 0.5 m, 1 m y 3 m. Estos efectos varían en relación a la distancia y son independientes de la longitud focal [2] . . . . . 2
1.2. Ejemplo de superposición craneofacial [21]. . . . . 4

2.1. Esquema de una red neuronal [14]. . . . . 6
2.2. Modelo neuronal para una neurona $k$ [11]. . . . . 7

# Índice de cuadros

# Capítulo 1

# Introducción

## 1.1. Definición del problema

La identificación facial ha adquirido gran relevancia durante la última década. La revolución del aprendizaje profundo y los sistemas automáticos de reconocimiento facial han llevado a una expansión del mercado desde los campos de la aplicación de la ley y la ciencia forense hasta áreas en el sector privado:comercio minorista, aplicaciones multimedia o seguridad. Además, el desarrollo de la tecnología de imagen ha mejorado tanto la calidad como la disponibilidad de datos fotográficos,lo que también ha contribuido a la aplicación de técnicas de identificación multimodal mediante el uso de modelos faciales en 3D o imágenes médicas [17, 25].

Las técnicas de identificación normalmente son realizadas por expertos con o sin la ayuda de sistemas automáticos. Los expertos analizan los datos y evalúan las características anatómicas de un individuo desconocido para compararlas con las de uno o varios individuos conocidos. Actualmente existen cuatro métodos de comparación facial reconocidos: análisis morfológico, superposición, foto-antropometría y comparación holística [7]. Para que este análisis sea confiable y concluyente, los datos (fotografías faciales), deben estar en unas condiciones adecuadas (calidad, resolución, enfoque o iluminación) y la escena (punto de vista de la cámara, pose de la cabeza, expresión facial) debe ser lo más neutral y representativa posible. Estos requisitos nos aseguran que los rasgos faciales sean más fieles a las características anatómicas del individuo, y por tanto, permiten que las técnicas de identificación sean más robustas.

Muchos estudios han identificado limitaciones en los actuales métodos de reconocimiento automático. Los principales factores más desafiantes son la pose, la iluminación, la expresión y la variación en la edad [13, 15]. Sin embargo también existen otros factores importantes como la oclusión, el

género o la etnia [5, 8].

Uno de los factores más determinantes es la distorsión de perspectiva [16] // que es una deformación o transformación de un objeto y su entorno que difiere significativamente de cómo se vería el objeto con una longitud focal normal, debido a la escala relativa de las características cercanas y distantes //. En nuestro caso, el objeto es el rostro de un individuo, por ejemplo, podríamos observar la deformación de los rasgos faciales (orejas, nariz o forma de la cara) debido a la cercanía de la cámara al sujeto durante la adquisición de la fotografía [24] (ver Figura 1.1). La distorsión de perspectiva tiene un efecto negativo en los sistemas de reconocimiento automático [6, 18, 23], lo que a su vez puede dificultar la identificación precisa de individuos. La distorsión de perspectiva está estrechamente relacionada con la distancia cámara-sujeto (subject-to-camera distance, SCD en adelante), de hecho, la relación es de decremento logarítmico, esto significa que, valores pequeños de SCD corresponden con una mayor distorsión, mientras que la distorsión disminuye conforme el SCD aumenta [20]. Por esta razón, este TFG trata sobre la estimación del SCD en fotografías faciales.

Figura 1.1: Efectos de la distorsión de perspectiva en características faciales de fotografías realizadas a diferentes SCD: 0.5 m, 1 m y 3 m. Estos efectos varían en relación a la distancia y son independientes de la longitud focal [2]

La estimación del SCD en imágenes faciales abre la posibilidad a cuantificar las diferencias en la distorsión entre dos conjuntos de imágenes, y además, la posibilidad de reproducir las condiciones originales de la escena cuando hay disponibles modelos faciales 3D o restos esqueléticos. Esta última característica se considera esencial tanto para técnicas de identificación manual como automáticas, ya que mejoran la credibilidad de las comparaciones faciales mediante la medición y control de una mayor fuente de incertidumbre.

El único método totalmente automatizado para estimar el SCD en fotografías faciales, hasta la fecha, se llama FacialSCDnet [2]. Este método

utiliza una arquitectura basada en deep learning (VGG-16) para procesar fotografías faciales y estimar el SCD. La arquitectura VGG-16 es ampliamente reconocida por sus habilidades para aprender tareas. Las capas convolucionales del modelo VGG-16, usadas en FacialSCDnet, son pre-entrenadas con el conjunto de datos ImageNet y se reajustan (fine-tuning) con un conjunto de datos diseñado específicamente para la tarea de estimar el SCD. Este conjunto de datos específico se compone de una parte sintética y una parte real.

El propósito de este TFG es mejorar el método actual del estado del arte en la estimación automática del SCD. Para ello, se plantea mejorar el proceso de generación de una base de datos sintética de manera que haya modelos 3D más completos (de cuerpo entero y poses distintas) con fondos e iluminación más realistas. Por otro lado, se explorarán mejoras adicionales como un cambio de framework que optimice los procesos de entrenamiento, el uso de un sistema de image augmentation que emplee GPU, o el desarrollo y uso de nuevas arquitecturas que mejoren los resultados.

## 1.2. Motivación

En el ámbito forense, el principal foco recae sobre la determinación de la identidad humana cuando existe información esquelética [10]. En las últimas décadas, los antropólogos han centrado su atención en mejorar las técnicas para realizar una identificación más precisa. En este contexto, la estimación del SCD juega un papel crucial, ya que, si estimamos el SCD con fiabilidad, podemos recrear la imagen con los restos esqueléticos (aplicando el valor del SCD). A continuación, realizamos las comparaciones anatómicas mediante la superposición craneofacial [22] para identificar si se corresponde con la misma persona (ver Figura 1.2). Si los parámetros de adquisición entre ambas imágenes (normal y esquelética) son distintos entonces dificultaría el análisis morfológico [21].

Por otra parte, sabemos que el SCD y la distorsión de perspectiva (PD) tienen una relación de decremento logarítmico [20]. Esto significa que, para cuantificar el nivel de distorsión entre dos imágenes, bastaría con estimar el SCD para cada una de ellas, y mediante la siguiente fórmula hayamos el porcentaje de distorsión:

$$PD(\,\%) = (\frac{B'}{A'} - 1) \times 100$$

donde A' y B' son las proyecciones, en el sensor de la cámara, de los objetos A y B.

- Corrección de imágenes ?

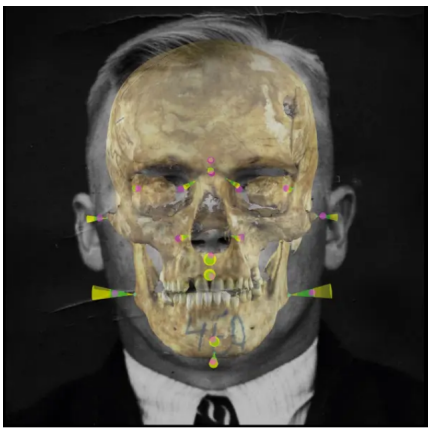

Figura 1.2: Ejemplo de superposición craneofacial [21].

## 1.3. Objetivos

El objetivo general de este Trabajo de Fin de Grado (TFG) consiste en desarrollar un método adecuado para abordar el problema de la estimación de la distancia cámara-sujeto (SCD) 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 para la estimación del SCD en fotografías faciales.
2. Generar un conjunto de datos sintético de mayor calidad (modelos 3D más completos de cuerpo entero y poses distintas, con fondos e iluminación más realistas)
3. Realizar un estudio experimental que permita validar los enfoques propuestos y extraer conclusiones sobre su aplicabilidad al problema.
4. Desarrollar y entrenar con nuevas arquitecturas que mejoren los resultados
5. Usar tecnologías más recientes que mejoren los tiempos de aprendizaje y los resultados obtenidos

# Capítulo 2

# Fundamentos teóricos

? Distancia cámara sujeto

? Distorsión de perspectiva

? Reconocimiento facial

? Transferencia de aprendizaje

## 2.1. Aprendizaje automático

El aprendizaje automático (Machine Learning, ML) [12, 4] es una rama de la IA 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, 19]:

El **aprendizaje supervisado** consiste en entrenar con datos de los que se saben sus etiquetas (una etiqueta nos indica qué es cada dato). Por ejemplo, los datos de entrada podrían ser imágenes de animales y sus etiquetas podrían ser "perro.º "gato". A partir de los datos y sus etiquetas, el agente aprende una función que dado un nuevo dato, predice su etiqueta. Es el tipo de aprendizaje más utilizado, los datos vienen ya 'preparados' para su uso. Es el tipo de aprendizaje que utilizaremos en este TFG.

En el **aprendizaje no supervisado** el agente aprende los patrones de los datos de entrada sin ninguna realimentación, es decir, los datos no están etiquetados. La herramienta más usada en el aprendizaje no supervisado es el agrupamiento, que consiste en detectar potenciales grupos en los datos de entrada. Este enfoque requiere un mayor número de datos.

En el **aprendizaje por refuerzo** el agente aprende mediante una serie de recompensas o castigos. El agente intentará realizar acciones que le

proporcionen mejores recompensas en el futuro. Este tipo de aprendizaje es muy utilizado para enseñar a jugar a juegos.

## 2.2. Aprendizaje profundo

### 2.2.1. Redes neuronales

Las redes neuronales (ANNs) [9, 11, 3] son redes computacionales que intentan, a groso modo, simular el proceso de decisión de las neuronas del sistema nervioso central de los animales o humanos. Las ANNs poseen unidades de procesamiento de información llamadas neuronas, que están conectadas entre sí mediante capas. La red se compone de (ver Figura 2.1):

- 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, que representa el valor o valores predichos

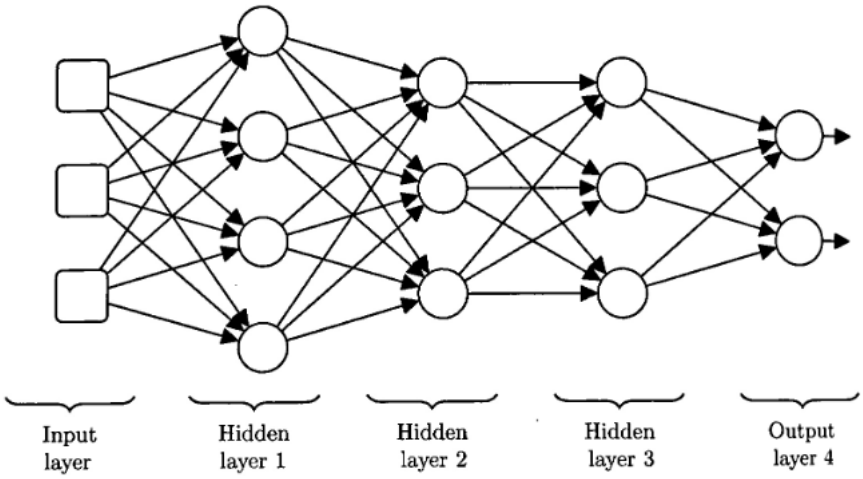


Figura 2.1: Esquema de una red neuronal [14].

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 2.2:

- 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 y la función ReLU.

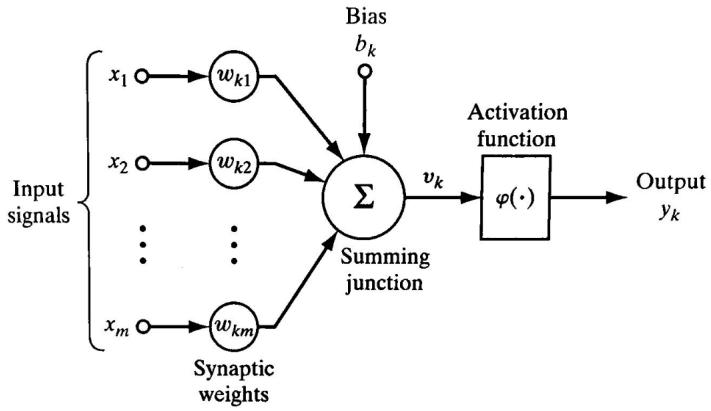


Figura 2.2: Modelo neuronal para una neurona $k$ [11].

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

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

siendo $\phi$ 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 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

#### Capa de pooling

# Capítulo 3

# Estado del Arte

## 3.1. Estimación de distancia cámara-sujeto

En el campo del aprendizaje automático, el tema de la estimación de la profundidad en fotografías ha ganado recientemente mucha atención.

¿Qué se ha hecho a lo largo del tiempo?

1. Van Dijk, T., De Croon, G. (2019). How do neural networks see depth in single images? In IEEE/CVF international conference on computer vision (pp. 2183–2191).

## 3.2. FacialSCDnet

# Capítulo 4

# Materiales y métodos

## 4.1. Materiales

## 4.2. Métodos

# Capítulo 5

# Experimentos

## 5.1. ...

## 5.2. ...

## 5.3. ...

# Capítulo 6

# Conclusiones y trabajos futuros

## 6.1. ...

## 6.2. ...

## 6.3. ...

# Bibliografía


[1] Yaser S. Abu-Mostafa, M. Magdon-Ismail y H.T. Lin. *Learning from Data: A Short Course*. AMLBook.com, 2012. ISBN: 978-1-60049-006-4. URL: https://books.google.co.uk/books?id=iZUzMwEACAAJ.

[2] Enrique Bermejo y col. «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. ISSN: 0957-4174. DOI: https://doi.org/10.1016/j.eswa.2022.118457.

[3] C.M. Bishop. *Neural networks for pattern recognition*. Oxford University Press, USA, 1995.

[4] Sara Brown. «Machine learning, explained». En: *Massachusetts Institute of Technology* (). URL: https://mitsloan.mit.edu/ideas-made-to-matter/machine-learning-explained.

[5] Joy Buolamwini y Timnit Gebru. «Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification». En: *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*. Ed. por Sorelle A. Friedler y Christo Wilson. Vol. 81. Proceedings of Machine Learning Research. PMLR, 23–24 Feb de 2018, págs. 77-91. URL: https://proceedings.mlr.press/v81/buolamwini18a.html.

[6] Naser Damer y col. «Deep Learning-based Face Recognition and the Robustness to Perspective Distortion». En: *2018 24th International Conference on Pattern Recognition (ICPR)*. 2018, págs. 3445-3450. DOI: 10.1109/ICPR.2018.8545037.

[7] FISWG. «Facial Comparison Overview and Methodology Guidelines V2.0». En: (2022). URL: https://fiswg.org/fiswg_facial_comparison_overview_and_methodology_guidelines_V2.0_2022.11.04.pdf.

[8] Nicholas Furl, P.Jonathon Phillips y Alice J O'Toole. «Face recognition algorithms and the other-race effect: computational mechanisms for a developmental contact hypothesis». En: *Cognitive Science* 26.6 (2002), págs. 797-815. ISSN: 0364-0213. DOI: https://doi.org/10.1016/S0364-0213(02)00084-8.

[9] Daniel Graupe. *Principles of artificial neural networks*. 3rd ed. World Scientific, 2007. ISBN: 9789814522731.

[10] «Handbook on Craniofacial Superimposition». En: ().

[11] Simon Haykin. *Neural Networks and Learning Machines*. 3rd ed. Pearson, 2009. ISBN: 9780131293762.

[12] IBM. *What is machine learning?* URL: https://www.ibm.com/topics/machine-learning.

[13] Anil K. Jain, Brendan Klare y Unsang Park. «Face Matching and Retrieval in Forensics Applications». En: *IEEE MultiMedia* 19.1 (2012), págs. 20-20. DOI: 10.1109/MMUL.2012.4.

[14] John D. Kelleher. *Deep learning*. The MIT Press, 2019. ISBN: 9780262537551.

[15] Zhifeng Li, Unsang Park y Anil K. Jain. «A Discriminative Model for Age Invariant Face Recognition». En: *IEEE Transactions on Information Forensics and Security* 6.3 (2011), págs. 1028-1037. DOI: 10.1109/TIFS.2011.2156787.

[16] *Perspective distortion*. URL: https://www.wikiwand.com/en/Perspective_distortion_(photography).

[17] Fred W. Prior y col. «Facial Recognition From Volume-Rendered Magnetic Resonance Imaging Data». En: *IEEE Transactions on Information Technology in Biomedicine* 13.1 (2009), págs. 5-9. DOI: 10.1109/TITB.2008.2003335.

[18] Zahid Riaz. y Michael Beetz. «On the effect of perspective distortions in face recognition». En: *Proceedings of the International Conference on Computer Vision Theory and Applications (VISIGRAPP 2012) - Volume 2: VISAPP*. INSTICC. SciTePress, 2012, págs. 718-722. ISBN: 978-989-8565-03-7. DOI: 10.5220/0003859107180722.

[19] Stuart Russell y Peter Norvig. *Artificial Intelligence: A Modern Approach 4ºedition*. Pearson, 2021. ISBN: 978-1-292-40113-3.

[20] Carl N. Stephan. «Perspective distortion in craniofacial superimposition: Logarithmic decay curves mapped mathematically and by practical experiment». En: *Forensic Science International* 257 (2015), 520.e1-520.e8. ISSN: 0379-0738. DOI: https://doi.org/10.1016/j.forsciint.2015.09.009.

[21] Carl N. Stephan. «Perspective distortion in craniofacial superimposition: Logarithmic decay curves mapped mathematically and by practical experiment». En: *Forensic Science International* 257 (2015), 520.e1-520.e8. ISSN: 0379-0738. DOI: https://doi.org/10.1016/j.forsciint.2015.09.009.

[22] Carl N. Stephan. «Estimating the Skull-to-Camera Distance from Facial Photographs for Craniofacial Superimposition». En: *Journal of Forensic Sciences* 62.4 (2017), págs. 850-860. DOI: https://doi.org/10.1111/1556-4029.13353.

[23] Joachim Valente y Stefano Soatto. «Perspective distortion modeling, learning and compensation». En: *2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. 2015, págs. 9-16. DOI: 10.1109/CVPRW.2015.7301314.

[24] Brittany Ward y col. «Nasal Distortion in Short-Distance Photographs: The Selfie Effect». En: *JAMA Facial Plastic Surgery* 20.4 (2018). PMID: 29494735, págs. 333-335. DOI: 10.1001/jamafacial.2018.0009.

[25] Mineo Yoshino y col. «Assessment of Computer-assisted Comparison between 3D and 2D Facial Images». En: *Japanese journal of science and technology for identification* 5.1 (2000), págs. 9-15. DOI: 10.3408/jasti.5.9.