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, 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 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

Índice de figuras

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 [24, 34].

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 [8]. 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 [15, 21]. Sin embargo también existen otros factores importantes como la oclusión, el

género o la etnia [5, 10].

Uno de los factores más determinantes es la distorsión de perspectiva [22] // 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 [32] (ver Figura 1.1). La distorsión de perspectiva tiene un efecto negativo en los sistemas de reconocimiento automático [7, 26, 31], 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 [29]. 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 la SCD en fotografías faciales, hasta la fecha, se llama FacialSCDnet [2]. Este método utiliza una arquitectura basada en deep learning (VGG-16) para procesar fo-

tografías 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.

. 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 [12]. 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 [30] 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 [29].

Por otra parte, sabemos que el SCD y la distorsión de perspectiva (PD) tienen una relación de decremento logarítmico [29]. 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.

- Mitigar los efectos de la distorsión DISCO

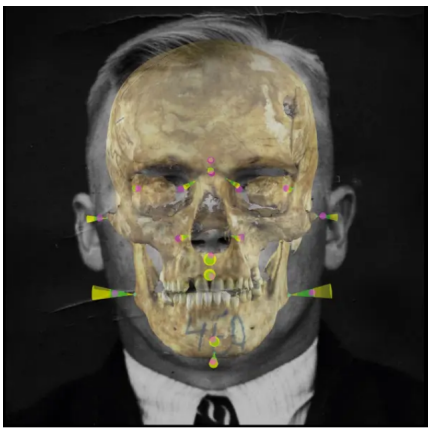

Figura 1.2: Ejemplo de superposición craneofacial [29].

1.3. Objetivos

El objetivo general de este Trabajo de Fin de Grado (TFG) consiste en mejorar el método actual del estado del arte en la estimación automática 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 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

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 [23], 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 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	Febrero				Marzo				Abril					Mayo				Junio		
	- Horas	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

Capítulo 2

Fundamentos teóricos

2.1. Aprendizaje automático

El aprendizaje automático (Machine Learning, ML) [14, 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, 27]:

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) [11, 13, 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

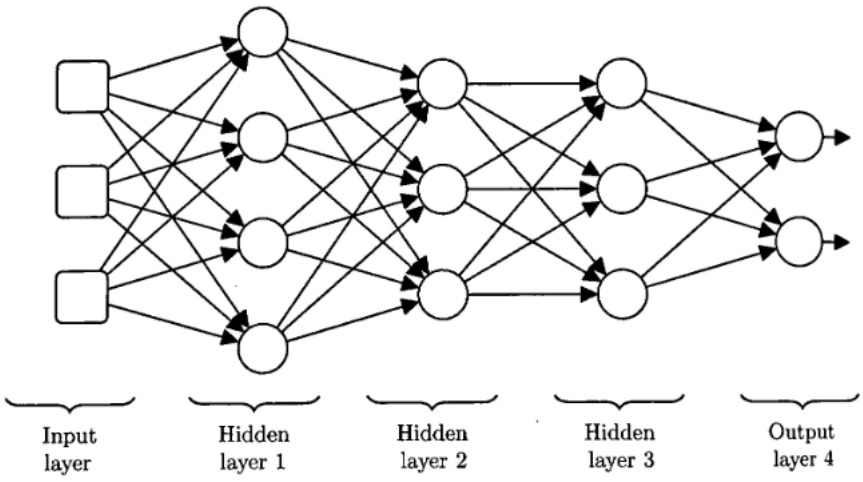

Figura 2.1: Esquema de una red neuronal [16].

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.

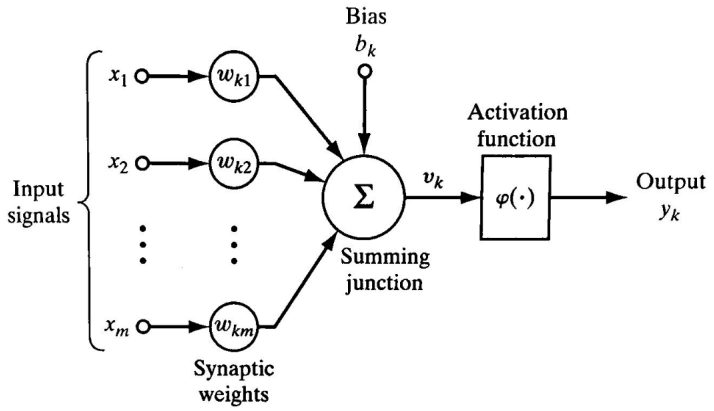

Figura 2.2: Modelo neuronal para una neurona k [13].

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 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 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 da-

tos 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 Network, CNN) [18, 19, 33] son un tipo de red neuronal profunda que trabaja con patrones de cuadrícula, como pueden ser imágenes.

A. Dhillon and G. Verma. Convolutional neural network: a review of models, methodologies and applications to object detection. Progress in Artificial Intelligence, 9(2):85–112, 2019

R. Yamashita, M. Nishio, R. K. G. Do y K. Togashi, ((Convolutional neural networks: an overview and application in radiology,)) Insights into Imaging, vol. 9, n.o 4, págs. 611-629, 2018.

Capa de convolución

Capa de pooling

Capa de normalización

Capa de activación

Capa densamente conectada

2.2.3. Transferencia de aprendizaje

2.3. Parámetros de la cámara y perspectiva

Distancia cámara sujeto

Longitud focal

Distorsión de perspectiva

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 3.1 la cantidad de publicaciones existentes en la base de datos Scopus [1] que hacen referencia a la estimación del SCD. Hay 430 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 deep learning. 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 del SCD fue propuesto por Flores et al. [9], se basa en el uso de un conjunto de puntos de referencia de la cara para estimar la posición y la distancia a la cámara en distancias desde 10 cm hasta 3 m. El proceso consiste en, entrenar un modelo de regresión para predecir la distancia entre la cámara y la cara, utilizando un conjunto de imágenes 3D de caras humanas con sus respectivos puntos de referencia faciales (se observa que estas referencias no varían drásticamente entre individuos, sino que se agrupan en 'clusters'). Dicho conjunto de datos 3D solo contiene vistas de frontales y de perfil 3/4.

Dada una imagen 2D de una cara desconocida, se identifican los puntos de referencia faciales y mediante el algoritmo EPnP [20], junto con la suposición de que las referencias no varían mucho entre individuos, se infiere

[1] Las búsquedas se pueden consultar en el apéndice

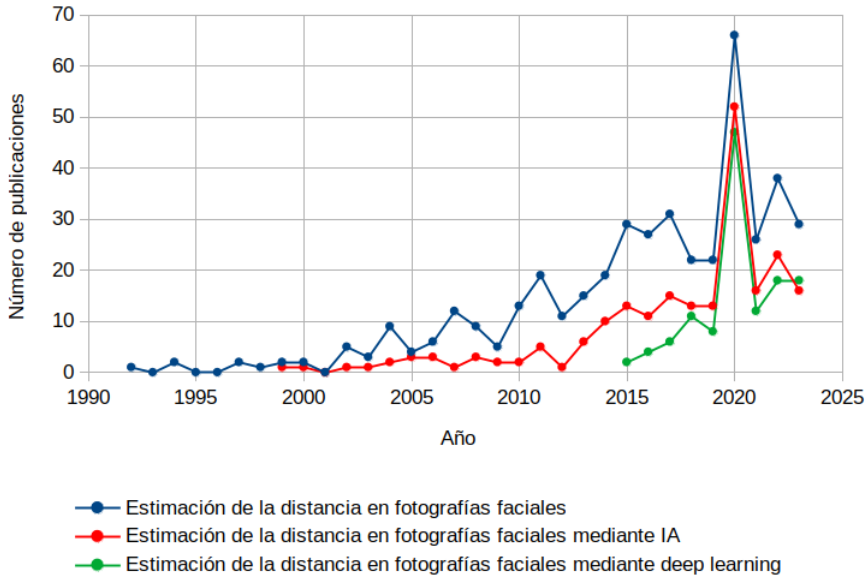

Figura 3.1: Número de publicaciones, en Scopus, relacionadas con la estimación de la distancia en fotografías faciales en función del año de publicación

la posición 3D de la cara y se utiliza el modelo de regresión previamente entrenado para predecir el valor de la distancia a la cámara.

Algunas de las limitaciones asociados a este primer método son: el uso de un conjunto de datos 3D (no siempre tendremos disponibles imágenes 3D para el entrenamiento), la combinación de diferentes longitudes focales en el mismo dataset o el reconocimiento manual de los puntos de referencia faciales

Posteriormente, en Burgos-Artizzu et al. [6] se propuso un método que no necesitaba la reconstrucción 3D ni la anotación manual de los puntos de referencia de la imagen. Este método utiliza un conjunto de datos llamado Caltech Multi-Distance Portraits (CMDP), compuesto de 53 retratos individuales desde 7 distancias distintas entre 60 cm y 480 cm, para entrenar el modelo. Todas las imágenes del conjunto de datos fueron anotadas manualmente con 55 marcas faciales.

El método se basa en dos fases: primero, la identificación automática de los puntos de referencia faciales, y después, la estimación de la distancia mediante regresión.

Este nuevo enfoque, a pesar de mejorar lo que previamente se había hecho, sigue teniendo algunas limitaciones como el recorte de las imágenes (pérdida de resolución) o la única vista frontal.

Existen otros métodos que estiman el SCD a partir de características

anatómicas como el tamaño de la cara [28], la distancia de los ojos [25] o una combinación de ambos [17].

3.2. MediaPipe Iris

MediaPipe Iris [2] es un modelo de aprendizaje automático, creado por investigadores de Google, capaz de seguir puntos de referencia (iris, pupila y contornos del ojo) usando una cámara RGB, en tiempo real, y sin necesidad de ningún hardware especializado. A través de los puntos de referencia del iris, el model es capaz de determinar la distancia métrica entre el sujeto y la cámara.

El modelo se basa en el diámetro horizontal del iris del ojo humano, que se mantiene relativamente constante en un rango de 11.7 ± 0.5 mm en una amplia población, esto junto con algunos simples argumentos geométricos, permiten la posibilidad de estimar el SCD.

Este modelo requiere de unas condiciones, que además, son sus principales limitaciones. El modelo solo puede ser usado cuando: existen datos EXIF disponibles; son imágenes frontales donde el iris es visible; y los individuos están a menos de 2m de la posición de la cámara.

3.3. PerspectiveX

PerspectiveX fue el método propuesto por Stephan et al. [30] para la estimación del SCD en fotografías faciales con el fin de mejorar el proceso de superposición craneofacial.

Este método se basa en la localización de una característica anatómica, la longitud de la fisura palpebral entre dos marcas precisas y fácilmente determinables. Se utiliza este rasgo anatómico debido a que: es fácilmente visible en vista frontal e incluso en el lado más cercano a la cámara si la cabeza está girada; está definida por dos puntos de referencia muy precisos; tiene una variación muy ajustada, debido a restricciones evolutivas; es una característica facial relativamente grande, por lo que, minimiza el impacto de los errores en comparación con otras características faciales invariantes como el diámetro del iris; tiene una distribución normal que reduce el error de predicción a la mitad.

Además de la fisura palpebral, PerspectiveX necesita conocer el tipo de cámara y la longitud focal de las lentes, que se pueden extraer siempre de imágenes electrónicas usando lectores EXIF disponibles en internet. El tipo de cámara es necesario para obtener las especificaciones de píxeles.

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

Finalmente, la estimación del SCD se realiza mediante la siguiente fórmula:

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

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)

Al no disponer de la longitud real de la fisura palpebral, se utiliza el valor medio de un conjunto de individuos agrupados por sexo y edad, ya que, sabemos que la longitud varía muy poco debido a restricciones evolutivas.

Este método permite una precisa estimación del SCD para una longitud focal conocida. Sin embargo, posee algunas limitaciones como: el requerimiento de interacción manual para anotar los puntos de referencia faciales; y que no tiene en cuenta las rotaciones de cabeza de más de 30^{o}.

3.4. FacialSCDnet

FacialSCDnet fue un método propuesto por Bermejo et al. [2], en el que se estima el SCD, directamente desde las fotografías, mediante la aplicación de técnicas deep learning. El uso de una arquitectura profunda evita una restricción crítica: el requerimiento de detectar una característica anatómica en particular para guiar el proceso de estimación. Es por ello que este método es capaz de estimar el SCD en cualquier posición de la cabeza, desde frontal hasta perfil lateral.

Se utilizó un conjunto de datos de entrenamiento compuesto por dos colecciones:

- Conjunto sintética: se generaron imágenes sintéticas 2D a partir de los modelos 3D de la base de datos Stirling ESRC 3D Face [3]. En particular, se usaron 315 modelos faciales de 54 indviduos diferentes para generar aproximadamente 150.000 fotografías sintéticas.

- Conjunto de fotografías digitales: se adquirieron fotografías de 28 individuos mediante el siguiente protocolo de adquisición: se consideraron 4 longitudes focales diferentes (27 mm, 35 mm, 50 mm, 85 mm) en formato completo; se utilizaron 12 distancias, de la cámara al sujeto, desde 50 cm hasta 6 m; se fotografiaron 7 posiciones diferentes de la

[3]Stirling ESRC 3D Face: https://pics.stir.ac.uk/ESRC/index.htm

cabeza, desde el perfil izquierdo hasta el perfil derecho, a intervalos de 30^{o} de rotación.

La función de pérdida utilizada en el modelo, se basa en el error absoluto medio de la distorsión facial relativa:

$$Distorsion = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n} \tag{3.2}$$

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):

$$D_f = \frac{1}{1 + \frac{SCD}{d}} \tag{3.3}$$

En la ecuación 3.3, $d = 12.6572cm$ corresponde a un valor derivado de cálculos geométricos [29] para obtener experimentalmente el factor de distorsión de una cabeza humana de tamaño promedio, según el SCD de la fotografía.

FacialSCDnet está compuesto por 4 modelos de deep learning, cada uno asociado a una longitud focal de las utilizadas en el conjunto de datos. La estructura de cada CNN se basa en una arquitectura VGG-16, cuyos pesos están inicializados a los pre-entrenados con ImageNet [4]. Para adaptar la arquitectura al problema de estimación del SCD, los 5 bloques convolucionales se mantienen, se elimina la capa superior, y se añaden 2 capas totalmente conectadas que se entrenarán desde cero. Finalmente, la última capa del modelo consistirá en una activación lineal que realiza la tarea de regresión.

El proceso de entrenamiento de los modelos tuvo dos fases. Primero, los modelos fueron entrenados con el conjunto de datos sintético para aprender relaciones entre SCD y características faciales. Después, se utiliza un ajuste fino usando el conjunto de datos real.

[4]ImageNet: https://www.image-net.org/

Capítulo 4

Materiales y métodos

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).

4.1. Materiales

4.1.1. Modelos 3D

En este apartado se describe cómo se obtuvieron los conjuntos de datos 3D y las decisiones tomadas para elegir unos dataset frente a otros. Además, se explicará el subconjunto de casos elegidos para generar las fotografías faciales 2D.

Modelos faciales

Aquí sobre los modelos HeadSpace, H3DS-net, ...

Modelos de cuerpo entero

Aquí sobre HuMMan, People Snapshot, ...

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+

Capítulo 5

Experimentos

5.1. Detalles técnicos de la implementación

5.1.1. Entorno de desarrollo

En este apartado describimos el entorno que vamos a usar para entrenar el modelo.

5.1.2. Obtención del conjunto de datos

En este apartado se explica cómo se obtienen las imágenes que posteriormente se van a usar para entrenar el modelo. Este apartado no explica cómo se generan (eso sería en el 4.2.1) sino cómo se organizan (según longitud focal y distancia) ya que tenemos que entrenar 4 modelos.

5.1.3. Entrenamiento del modelo

En este apartado se describe qué hay que hacer para entrenar el modelo (ejecutar ciertos archivos) y en qué sistemas se va a entrenar (GPU de la UGR mediante SSH, etc)

5.2. Experimentos

5.2.1. Protocolo de validación experimental

Este apartado describimos el esquema de división del dataset en entrenamiento, validación y test.

5.2.2. Métricas

Este apartado explica las métricas que vamos a utilizar para medir el rendimiento de los modelos.

5.2.3. Experimentos con VGG16

En este apartado realizamos el entrenamiento de nuestros modelos con el nuevo dataset. Comparamos con los resultados de FacialSCDnet para ver si nuestro nuevo dataset mejora de algún modo el entrenamiento.

Además, en otro apartado se podrían utilizar otras arquitecturas distintas a VGG16 para comparar con esta.

Capítulo 6

Conclusiones y trabajos futuros

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.

[2] 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. 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*. Vol. 81. PMLR, 2018, págs. 77-91. URL: https://proceedings.mlr.press/v81/buolamwini18a.html.

[6] Xavier P. Burgos-Artizzu, Matteo Ruggero Ronchi y Pietro Perona. *Distance Estimation of an Unknown Person from a Portrait*. Springer International Publishing, 2014, págs. 313-327. ISBN: 978-3-319-10590-1. DOI: http://dx.doi.org/10.1007/978-3-319-10590-1_21.

[7] Naser Damer et al. «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: http://dx.doi.org/10.1109/ICPR.2018.8545037.

[8] 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.

[9] Arturo Flores et al. *Camera Distance from Face Images*. Springer Berlin Heidelberg, 2013, págs. 513-522. ISBN: 978-3-642-41939-3. DOI: http://dx.doi.org/10.1007/978-3-642-41939-3_50.

[10] 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.

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

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

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

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

[15] 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: http://dx.doi.org/10.1109/MMUL.2012.4.

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

[17] M.S. Shashi Kumar, K.S. Vimala y N. Avinash. «Face distance estimation from a monocular camera». En: *2013 IEEE International Conference on Image Processing*. 2013, págs. 3532-3536. DOI: http://dx.doi.org/10.1109/ICIP.2013.6738729.

[18] Y. LeCun et al. «Backpropagation Applied to Handwritten Zip Code Recognition». En: *Neural Computation* 1 (1989), págs. 541-551.

[19] Y. Lecun et al. «Gradient-based learning applied to document recognition». En: *Proceedings of the IEEE* 86.11 (1998), págs. 2278-2324. DOI: 10.1109/5.726791.

[20] 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). DOI: http://dx.doi.org/10.1007/s11263-008-0152-6.

[21] 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: http://dx.doi.org/10.1109/TIFS.2011.2156787.

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

[23] R.S. Pressman. *Software Engineering: A Practitioner's Approach*. McGraw-Hill higher education. Boston, 2005. ISBN: 9780073019338. URL: https://books.google.es/books?id=bL7QZHtWvaUC.

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

[25] Khandaker Abir Rahman et al. «Person to Camera Distance Measurement Based on Eye-Distance». En: *2009 Third International Conference on Multimedia and Ubiquitous Engineering*. 2009, págs. 137-141. DOI: http://dx.doi.org/10.1109/MUE.2009.34.

[26] 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*. SciTePress, 2012, págs. 718-722. DOI: http://dx.doi.org/10.5220/0003859107180722.

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

[28] Mohamed Tahir Ahmed Shoani, Shamsudin H. M. Amin e Ibrahim M. H. Sanhoury. «Determining subject distance based on face size». En: *2015 10th Asian Control Conference (ASCC)*. 2015, págs. 1-6. DOI: http://dx.doi.org/10.1109/ASCC.2015.7244491.

[29] 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.

[30] 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.

[31] 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: http://dx.doi.org/10.1109/CVPRW.2015.7301314.

[32] Brittany Ward et al. «Nasal Distortion in Short-Distance Photographs: The Selfie Effect». En: *JAMA Facial Plastic Surgery* 20.4 (2018), págs. 333-335. DOI: http://dx.doi.org/10.1001/jamafacial.2018.0009.

[33] 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. ISSN: 0167-8655. DOI: https://doi.org/10.1016/j.patrec.2018.05.018.

[34] Mineo Yoshino et al. «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: `http://dx.doi.org/10.3408/jasti.5.9`.

Búsquedas en Scopus

Estimación de la distancia en fotografías faciales (430)

TITLE-ABS-KEY ((distance OR depth) AND estimation AND (photographs OR images) AND facial) AND (LIMIT-TO (SUBJAREA , “COMP”) OR LIMIT-TO (SUBJAREA , “ENGI”))

Estimación de la distancia en fotografías faciales mediante IA (218)

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 , “COMP”) OR LIMIT-TO (SUBJAREA , “ENGI”))

Estimación de la distancia en fotografías faciales mediante deep learning (126)

TITLE-ABS-KEY ((deep AND learning) AND ((distance OR depth) AND estimation AND (photographs OR images) AND facial)) AND (LIMIT-TO (SUBJAREA , “COMP”) OR LIMIT-TO (SUBJAREA , “ENGI”))