Jornadas de Automática

Reconocimiento facial en informativos televisivos mediante redes convolucionales profundas

Asensi-González, Ricardo[a], Herrera, Pedro Javier[a,*]

[a] *Dpto. de Ingeniería de Software y Sistemas Informáticos, Universidad Nacional de Educación a Distancia, C/ Juan del Rosal, 16, 28040, Madrid, España.*

To cite this article: Asensi-González, Ricardo, Herrera, Pedro Javier. 2025. Face recognition in television news images using deep convolutional networks. Jornadas de Automática, 46.
https://doi.org/10.17979/ja-cea.2025.46.12046

Resumen

Este trabajo propone un sistema de inteligencia artificial basado en redes neuronales profundas que permite la detección y reconocimiento de personas concretas en imágenes extraídas de informativos televisivos. Para ello, se ha creado un conjunto de datos (*dataset*) que consta de 12800 imágenes, centrado principalmente en figuras políticas de ámbito nacional. El sistema propuesto realiza la detección del individuo en la escena de manera automática utilizando la red YOLOv8 y, posteriormente, realiza su reconocimiento a partir del clasificador que proporcione mayor certidumbre. Para ello, se compararon siete arquitecturas de red neuronal convenientemente adaptadas a esta problemática concreta: VGG-16, VGG-19, InceptionV3, Xception, ResNet-101, MobileNetV2 y DenseNet-169, siendo este último el modelo que obtiene en promedio un mejor desempeño en todas las pruebas realizadas. Los resultados confirman la viabilidad del sistema y permiten sentar las bases para futuras investigaciones.

Palabras clave: Procesamiento de imágenes, Redes neuronales, Aprendizaje máquina, Técnicas de inteligencia artificial, Visión por computador.

Face recognition in television news images using deep convolutional networks

Abstract

This work proposes an artificial intelligence system based on deep neural networks that enables the detection and recognition of specific persons in images extracted from television news. To this end, a dataset consisting of 12800 images was created, focusing primarily on Spanish political figures. The proposed system automatically detects the individual in the scene using the YOLOv8 network and subsequently recognizes the individual using the classifier that provides the greatest certainty. To this end, seven neural network architectures appropriately adapted to this specific problem were compared: VGG-16, VGG-19, InceptionV3, Xception, ResNet-101, MobileNetV2, and DenseNet-169, with the latter model achieving the best average performance across all tests. The results confirm the viability of the system and lay the groundwork for future research.

Keywords: Image processing, Neural networks, Machine learning, Artificial intelligence techniques, Computer vision.

1. Introducción

Disponer de un sistema automático capaz de reconocer personas en imágenes puede mejorar significativamente el proceso de documentación, especialmente en medios de comunicación como los centros de televisión. Este sistema facilitaría identificar, clasificar y almacenar imágenes para posteriormente recuperarlas de su archivo histórico, añadiendo valor a futuras noticias. El departamento de documentación en una televisión se encarga de la organización de grandes volúmenes de contenido audiovisual, como el archivo histórico de Radio Televisión Española (RTVE), que acumula cientos de miles de horas de material audiovisual. Tradicionalmente, este trabajo se realizaba de forma manual mediante descripción textual y generación de metadatos.

*Autor para correspondencia: pjherrera@issi.uned.es

Para poder alcanzar el objetivo de reconocer e identificar correctamente a un individuo concreto, se deben alcanzar una serie de subobjetivos o pasos intermedios:

1. Obtener el conjunto de imágenes de donde se extraerán los rostros para ser entrenados. En este trabajo se han utilizado las imágenes de los informativos emitidos por internet de la televisión pública de ámbito nacional RTVE, centrándose en personas dedicadas a la actualidad política.

2. Detectar las imágenes que contienen una persona o rostro humano. Para ello, en esta propuesta se ha desarrollado un programa informático específico, capaz de extraer las imágenes que presentan esas características.

3. Seleccionar, separar y clasificar las personas objeto de interés para su posterior identificación. Estas imágenes constituirán el conjunto de datos (*dataset*) sobre los que se realizará el proceso de aprendizaje del sistema desarrollado. El *dataset* creado como parte de la investigación desarrollada en este trabajo consta de 8 clases y 12800 imágenes.

4. Realizar un estudio comparativo de las principales arquitecturas existentes en la literatura para la detección de rostros. En este trabajo se identificaron inicialmente 21 modelos de red de los que finalmente se seleccionaron 7 que han sido convenientemente adaptados a este problema concreto y al *dataset* disponible.

El resto del documento se organiza de la siguiente manera. La segunda sección describe brevemente el estado del arte en la detección de objetos en general y la detección de rostros en particular, centrándose en los modelos basados en redes neuronales profundas, de gran aplicación en la actualidad. La tercera sección presenta el enfoque propuesto y describe el *dataset* utilizado. La cuarta sección muestra los resultados obtenidos a partir de la comparativa realizada, que cubre un número significativo de modelos empleados con éxito en la literatura para abordar esta misma problemática. Finalmente, la quinta sección presenta las principales conclusiones extraídas, así como las líneas de investigación abiertas.

2. Estado del arte

Los primeros trabajos relativos a la detección de rostros datan de la década de 1960, con un sistema capaz de clasificar de manera manual imágenes de rostros utilizando lo que se conoce como una tableta RAND (Bledsoe, 1963, 1964, 1966). En la década de 1970 se detallan características faciales que introducen mejoras en la precisión del reconocimiento facial, aunque la biometría aún es calculada manualmente (Goldstein *et al*, 1971). A finales de 1980, se comienza a aplicar el álgebra lineal a este problema a través de la técnica conocida como EigenFaces (Sirovich and Kirby, 1987), que codifica una imagen facial de tal forma que puede ser representada como la combinación lineal de otras imágenes. Se demuestra que con menos de cien valores es posible codificar con precisión una imagen de rostro normalizada. En la década siguiente se extiende la tecnología de EigenFaces, permitiendo detectar un rostro humano dentro de una fotografía, abriendo la puerta al reconocimiento facial automático (Turk and Pentland, 1991). Al inicio del siglo XXI se produce un gran avance gracias al algoritmo Viola-Jones, cuyo bajo coste computacional y gran rapidez, permite la detección de rostros en tiempo real de forma robusta y precisa (Viola and Jones, 2001). A partir de ese momento, tanto en la detección de objetos en general como en la detección de rostros en particular, se produce un aumento significativo de investigaciones y métodos basados en el uso de redes neuronales artificiales: multicapa, convolucionales, recurrentes o redes de base radial, entre otras. Dentro de las utilizadas para la detección de objetos destacan las redes convolucionales: Neocognitron (Fukushima, 1980), LeNet (LeCun *et al*., 1990), AlexNet (Krizhevsky *et al*., 2012), ZFNET (Zeiler and Fergus; 2014), VGG-16 y VGG19 (Simonyan and Zisserman, 2014) empleadas en diversas arquitecturas posteriores, como Faster R-CNN (Ren *et al*., 2015), Inception (Szegedy *et al*., 2014), ResNet (He *et al*., 2015), XCeption (Chollet, 2017), DenseNet (Huang *et al*., 2017), YOLOv8 (Jocher *et al*., 2023); redes con propuestas de regiones: R-CNN (Girshick *et al*., 2013), Fast-CNN (Girshick, 2014), Faster R-CNN (Ren *et al*., 2015) y Mask R-CNN (He *et al*., 2017); y redes específicas para detección de rostros: FaceNet (Schroff *et al*., 2015), MobileNets (Howard *et al*., 2017), MobileFaceNets (Chen *et al*., 2018), PyramidBox (Tang *et al*., 2018), DSFD (Li *et al*., 2019) y MixFaceNets (Boutrus *et al*., 2021), con aplicaciones, por ejemplo, en sistemas de vigilancia, de acceso a edificios o lugares controlados (Pajares *et al*., 2021).

Generalmente los diferentes modelos suelen ser entrenados con distintos conjuntos de datos, por lo que los resultados obtenidos varían sustancialmente según el conjunto de datos utilizado, dificultando una comparación directa entre todos ellos. El primer conjunto de datos destacable fue ImageNet (Deng *et al*., 2009), si bien en la actualidad existen un gran número de ellos como, por ejemplo, PASCAL VOC 2007 (Everingham *et al*., 2010), PASCAL VOC 2012 (Everingham *et al*., 2015), MS COCO (Lin *et al*., 2015), LFW (Huang et al., 2007), Youtube Faces DB (Wolf *et al*., 2011), MegaFaces (Nech and Kemelmacher-Shlizerman, 2017) y MS-Celeb-1M (Guo *et al*., 2016).

Se hace difícil, por tanto, poder resaltar a priori un modelo concreto como el más idóneo para abordar el problema específico que se plantea en el presente trabajo, a menos que se realice una comparación empírica que permita destacar, a partir del conjunto de datos disponible, el modelo que mejor se comporta. Es por ello por lo que se ha optado por realizar una comparativa entre los principales modelos existentes en la literatura, convenientemente adaptados a este problema concreto. Tras una primera selección de 21 modelos de red (Asensi-González, 2024), se escogieron finalmente los siete siguientes en base a su buen desempeño con el conjunto de datos ImageNet: VGG-16 y VGG-19 (Simonyan and Zisserman, 2014), Inception (Szegedy *et al*., 2014), ResNet (He *et al*., 2015), XCeption (Chollet, 2017), DenseNet (Huang *et al*., 2017) y MobileNets (Howard *et al*., 2017).

3. Solución propuesta

3.1. Dataset

Para la creación del *dataset* se emplearon las imágenes de los informativos emitidos por internet de la televisión pública de ámbito nacional RTVE, centrándose en personas del ámbito político. Dicha elección se justifica por la mayor

disponibilidad de imágenes de estas figuras públicas, al ocupar la actualidad política un espacio importante dentro de los informativos. En concreto, se han empleado 279 informativos (TD-1, 15:00 horas), desde el 01/09/2022 hasta el 11/08/2023, de los que se ha extraído cada fotograma mediante un programa software diseñado específicamente para la realización de este trabajo, con un diezmado de 5:1 para garantizar un equilibrio entre el movimiento de los gestos y expresiones faciales y la obtención de un número de imágenes suficientes en planos de corta duración. Se ha llevado a cabo un etiquetado manual de las imágenes con el objetivo de seleccionar únicamente a los individuos que integran el conjunto de datos. Para este propósito, se utilizó una red YOLOv8 debidamente adaptada, lo que permitió recortar de manera independiente cada uno de los individuos presentes en las imágenes. Posteriormente, se descartaron los que no pertenecen a ninguna de las clases contempladas por el sistema propuesto.

El *dataset* consta de 8 clases: "*Alberto Nuñez Feijóo*", "*Joe Biden*", "*Cuca Gamarra*", "*Ione Belarra*", "*Pedro Sánchez*", "*Yolanda Diaz*", "*Desconocido*" y "*NoFace*". La clase "*Desconocido*" aglutina los rostros identificados que no corresponden a ninguna de las clases anteriores, mientras que la clase "*NoFace*" aglutina posibles falsos positivos para su posterior tratamiento, tal y como se describirá en la sección siguiente. Cada clase consta de 1600 imágenes, de modo que el *dataset* consta en total de 12800. La Figura 1 muestra una selección de imágenes de entrenamiento de la clase "*Joe Biden*".

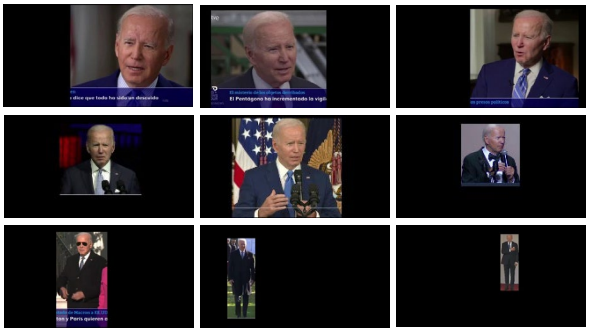

Figura 1: Muestras de entrenamiento de la clase "*Joe Biden*".

3.2. *Entrenamiento*

Para llevar a cabo el entrenamiento sobre el conjunto de datos, se selecciona cada uno de los modelos de red objeto de estudio (VGG-16, VGG-19, Inception, ResNet, XCeption, DenseNet y MobileNets). Sobre cada uno de ellos se han realizado numerosas pruebas variando los parámetros de configuración: función de pérdida, optimizador, número de épocas, pasos por época y pasos de validación. Por otro lado, y para aprovechar las ventajas que ofrece la técnica conocida como de ajuste fino (*fine-tuning*), se han "descongelado" (*unfrezing*) algunas de las últimas capas para ser entrenadas con nuestro conjunto de datos y así, mantener en las primeras capas las características extraídas de los datos originales con las que fueron entrenadas cada una de las redes. Por último, se añade un clasificador, con diferentes configuraciones tanto en número de capas como de neuronas en cada una de ellas, y se ha empleado la función *softmax* para realizar la clasificación. Todo ello con el objetivo de conseguir un incremento en la precisión alcanzada. La Figura 2 muestra un esquema con el diseño realizado para la fase de entrenamiento.

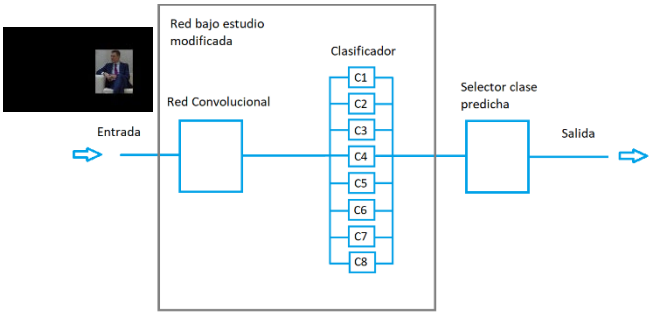

Figura 2: Esquema de la fase de entrenamiento.

3.3. *Clasificación*

Una vez entrenada la red se empleará el modelo generado para realizar inferencias sobre el conjunto de prueba o test. Para ello, como primera etapa de la arquitectura diseñada en este trabajo, se selecciona de manera automática en la imagen la localización de los *bounding boxes* correspondientes a los distintos individuos presentes en la escena (según el caso), agrupadas en una clase genérica denominada "*Persona*". Para ello, se ha utilizado la red YOLOv8 (Jocher *et al*., 2023) por su velocidad de ejecución y alta precisión. Para evitar o reducir los posibles falsos positivos cometidos por YOLO, se ha creado la clase denominada "*NoFace*" que permite detectar este error y tratarlo convenientemente. Los diferentes *bounding boxes* correspondientes a la clase "*Persona*", junto a sus imágenes recortadas, son procesados en la etapa siguiente de forma secuencial, introduciéndose como entrada a cada una de las arquitecturas de red seleccionadas para clasificación. La Figura 3 muestra el esquema de la arquitectura diseñada para realizar la inferencia con el conjunto de prueba o test.

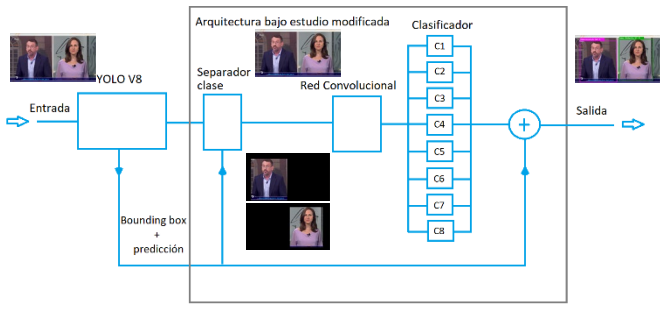

Figura 3: Esquema de la fase de clasificación.

4. **Análisis y evaluación de resultados**

Las pruebas se han dividido en dos grupos, en el primer grupo se han incluido 7 clases, descartando la clase "*Desconocido*", mientras que para el segundo grupo sí se han empleado las 8 clases establecidas. Se ha realizado esta separación para poder observar la influencia de la clase "*Desconocido*" en los resultados, debido a su especial

característica, donde cada muestra de la clase corresponde a una persona distinta, por lo que presenta una gran diversidad de características.

Para todas las arquitecturas seleccionadas se han realizado numerosas pruebas variando sus parámetros de configuración: función de pérdida, optimizador, número de épocas, pasos por época y pasos de validación. En la Tabla 1 se muestran los rangos de valores que se han empleado.

Tabla 1: Rango de ajuste de parámetros empleados en el entrenamiento

Optimizador	*SGD, RMSProp, Adam, Adagrad, Nadam, Adamax*
Tasa de aprendizaje	$[1\times10^{-1}, 1\times10^{-7}]$
Pasos por época	[12, 512]
Pasos de validación	[12, 264]
Épocas	[1, 1500]

Dadas las características para las que el sistema ha sido ideado, esto es, colaborar y facilitar la labor de documentación en un medio de comunicación como pudiera ser un centro de televisión, se debe seleccionar la arquitectura que mejor se adapte al tipo y necesidades concretas del departamento de documentación. En general, existen varios criterios como la velocidad, la precisión, el consumo de recursos y la disponibilidad del servicio. Sin embargo, el criterio principal que determinará la opción más idónea será la precisión obtenida por el sistema, ya que, en principio, al tratarse de analizar las imágenes que pasarán a formar parte del archivo no hay necesidad de tiempo real. En cambio, sí conviene que la clasificación sea lo más precisa posible.

4.1. Resultados obtenidos para el conjunto de 7 clases

Como se ha indicado anteriormente, para esta primera aproximación se ha considerado un subconjunto de las clases seleccionadas en esta investigación, ya que el objetivo de este punto es analizar cómo influye en general la no inclusión de la clase "*Desconocido*". Se han utilizado 7000 muestras para entrenamiento (1000 pertenecientes a cada clase), 2100 para validación (300 corresponden a cada clase) y 2100 para test (300 por cada clase). En la Tabla 2 se muestran los resultados obtenidos con cada una de las arquitecturas bajo estudio. El mejor resultado se consigue con la arquitectura DenseNet-169, alcanzando una precisión en la clasificación de 0,942 para el conjunto de 7 clases. Todas las arquitecturas analizadas, a excepción de VGG-19, logran una precisión por encima de 0,83. La Figura 4 muestra las gráficas de precisión y pérdida correspondientes a DenseNet-169.

Tabla 2: Precisión obtenida con 7 clases

Clase	VGG-16	VGG-19	Inception V3	Xception	DenseNet-169
A.N. Feijóo	0,786	0,587	0,803	0,826	0,897
J. Biden	0,803	0,770	0,850	0,779	0,944
C. Gamarra	0,819	0,604	0,755	0,704	0,911
I. Belarra	0,974	0,828	0,853	0,930	0,955
P. Sánchez	0,637	0,547	0,610	0,803	0,933
Y. Díaz	0,882	0,853	0,958	0,790	0,954
NoFace	0,990	1	1	0,997	1
Media	**0,842**	**0,741**	**0,833**	**0,833**	**0,942**

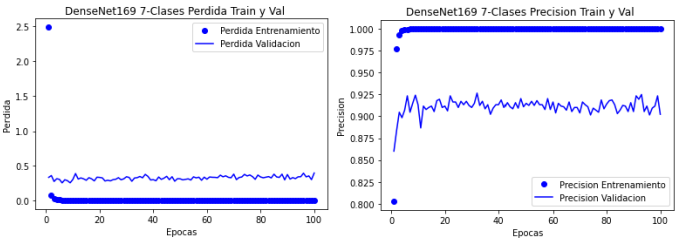

Figura 4: Gráficas de Pérdida (izq.) y Precisión (dcha.) para 7 clases con DenseNet-169.

4.2. Resultados obtenidos para el conjunto de 8 clases

En este segundo grupo de pruebas se han empleado todas las clases del conjunto de imágenes obtenidas para esta investigación, con el fin de analizar la influencia de la clase "*Desconocido*" en los resultados. Se han utilizado 8000 muestras para entrenamiento (1000 pertenecientes a cada clase), 2400 para validación (300 corresponden a cada clase) y 2400 para test (300 por cada clase). La Tabla 3 muestra los resultados obtenidos donde, en términos generales, se observa un ligero descenso en la precisión de entre 2 y 5 puntos porcentuales, salvo para VGG-16 que alcanza 6 puntos. Dicho descenso se justifica por la diversidad de características que contienen las muestras pertenecientes a la clase "*Desconocido*", compuesta por 1000 personas diferentes. No obstante, se alcanza una precisión máxima de 0,892 con el modelo DenseNet-169, que se posiciona como la mejor opción de entre las analizadas. Por el contrario, la arquitectura ResNet-101 es la que presenta peores resultados, con una precisión de 0,535. En la Figura 5 se muestran las gráficas de precisión y pérdida correspondientes a DenseNet-169, el modelo que mejores tasas de acierto ofrece. Con 8 clases los mejores resultados se alcanzaron después de 15 épocas, a diferencia del primer grupo, en el que los mejores resultados se alcanzaron después de 100 épocas (Figura 4).

En la Figura 6 se muestra la matriz de confusión, donde se observa que la mayor cantidad de falsos positivos ocurre entre las clases *"Cuca Gamarra"*, *"Pedro Sánchez"* y *"Yolanda Díaz"* con respecto a la clase *"Desconocido"*, alcanzando una tasa de falsos positivos entre 0,12 y 0,13. Entre las clases *"Yolanda Díaz"* y *"Cuca Gamarra"* la tasa de falsos negativos es de 0,08.

Tabla 3: Precisión obtenida con 8 clases

Clase	VGG-16	VGG-19	Inception V3	Xception	ResNet-101	Mobile Net V2	DenseNet-169
A.N. Feijóo	0,762	0,383	0,773	0,786	0,447	0,791	0,862
J. Biden	0,864	0,798	0,826	0,770	0,437	0,728	0,920
C. Gamarra	0,717	0,599	0,673	0,621	0,434	0,822	0,795
I. Belarra	0,834	0,777	0,898	0,904	0,662	0,796	0,974
P. Sánchez	0,463	0,607	0,600	0,677	0,093	0,703	0,817
Y. Díaz	0,748	0,844	0,769	0,639	0,559	0,769	0,794
Desconocido	0,867	0,750	0,853	0,867	0,690	0,860	0,970
NoFace	0,987	0,980	1	1	0,957	0,983	1
Media	**0,780**	**0,717**	**0,799**	**0,783**	**0,535**	**0,807**	**0,892**

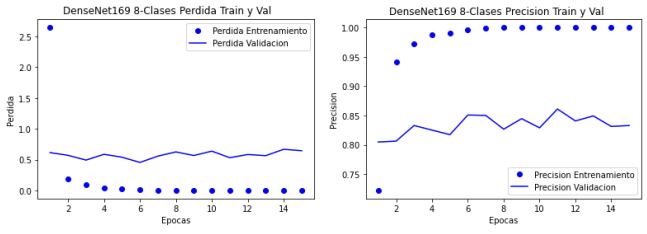

Figura 5: Gráficas de Pérdida (izq.) y Precisión (dcha.) para 8 clases con DenseNet-169.

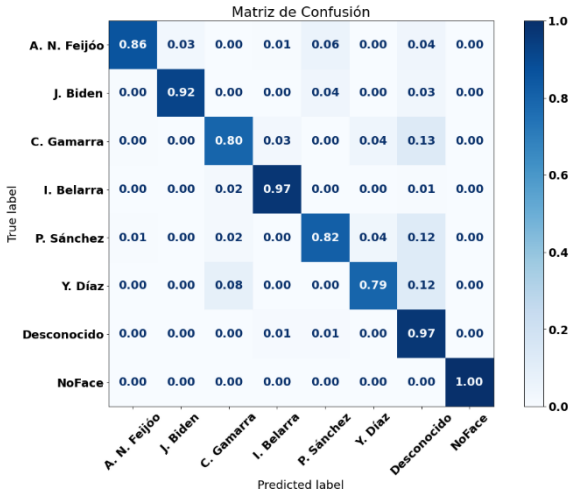

Figura 6. Matriz de confusión.

Por último, y a efectos ilustrativos, se muestran los resultados obtenidos con el enfoque propuesto, en combinación con el modelo DenseNet-169, Figura 7.

Figura 7. Ejemplos de predicción de clase.

5. Conclusiones

La detección de rostros es un caso particular de la detección de objetos, una tarea crítica que implica identificar y localizar con precisión ciertos elementos dentro de imágenes o secuencias de video. Además, el rostro humano puede presentar características muy similares entre un individuo y otro, por ejemplo, a nivel de ojos, nariz, boca, etc., lo que aumenta la complejidad del proceso de reconocimiento dadas las ligeras diferencias.

A lo largo de los años, se han ido desarrollando una serie de algoritmos para hacer frente a este desafío, dando lugar progresivamente a nuevos avances y mejoras. Este trabajo ha realizado un análisis de los principales modelos y arquitecturas desarrolladas para este fin. En concreto, y tras el pertinente estudio del estado del arte, se realizó una preselección y análisis de 21 modelos que aglutinan redes convolucionales, redes con propuestas de regiones y redes específicas para detección de rostros. Y finalmente se escogieron siete: VGG-16, VGG-19, InceptionV3, Xception, ResNet-101, MobileNetV2 y DenseNet-169, que han sido convenientemente adaptados y combinados apoyándose además en la red YOLOv8 para el procesamiento de la imagen previo a la clasificación.

Por otro lado, se ha elaborado un *dataset* propio que consta de 12800 imágenes y 8 clases. Se ha realizado una comparativa de los resultados obtenidos con los diferentes modelos en términos de precisión, identificando así la arquitectura que obtiene mejores resultados para el conjunto de datos específico utilizado en este trabajo. En concreto, el modelo que ha obtenido una mayor precisión es DenseNet-169 con 0,892 que se incrementa hasta 0,942 si no se tienen en cuenta los rostros identificados como "*Desconocido*", y que no corresponden a ninguna de las clases que soporta actualmente el sistema.

Precisamente, como líneas de trabajo futuro se plantea la extensión del *dataset*, incorporando un mayor número de clases. También se propone incorporar un módulo de reconocimiento de voz que, a partir de las declaraciones que suelen acompañar su presencia en los informativos, ayude en la identificación del individuo. Igualmente, se plantea la incorporación de un módulo de reconocimiento de caracteres que, a partir de los rótulos que suelen incluirse en la parte inferior de las imágenes, proporcione un criterio adicional para la identificación. Por último, se propone extender el sistema a partir del análisis del movimiento de los gestos, expresiones y su ritmo en las secuencias de vídeo que, además de establecer una serie de características que podrían ayudar también a la identificación del individuo, permitiría determinar si la persona que aparece en la imagen es real o ha sido generada de manera artificial (*deepfake*) utilizando, por ejemplo, redes generativas.

Agradecimientos

Este trabajo ha sido realizado gracias al apoyo de los proyectos DARWEEM (PID2020-113229RB-C43) y RECOVERY (PID2020-112658RB-I00), financiados por la

Agencia Estatal de Investigación (AEI/10.13039/501100011033).

Referencias

Asensi-González, R., 2024. Reconocimiento del rostro humano en imágenes de informativos televisivos mediante redes convolucionales profundas, Trabajo de Fin de Máster en Investigación en Ingeniería de Software y Sistemas Informáticos, Universidad Nacional de Educación a Distancia, Madrid.

Bledsoe, W. W., 1963. A study to determine the feasibility of a simplified face recognition machine. Panoramic Research, Inc. Palo Alto, California.

Bledsoe, W. W., 1964. Facial recognition project. Panoramic research, Inc. Palo Alto, California.

Bledsoe, W. W., 1966. Man-machine facial recognition: report on a large-scale experiment. Technical Report PRI 22, Panoramic Research, Inc. Palo Alto, California.

Boutrus, F., Damer, N., Fang, M., Kirchbuchner, F. Kuijper, A., 2021. MixFaceNets: extremely efficient face recognition networks. IEEE International Joint Conference on Biometrics (IJCB), pp. 1-8. DOI: 10.1109/IJCB52358.2021.9484374

Chen, S., Liu, Y., Gao, X, Han, Z., 2018. MobileFaceNets: efficient CNNs for accurate real-time face verification on mobile devices. In: Zhou, J., et al. Biometric Recognition. CCBR 2018. Lecture Notes in Computer Science. Vol. 10996. Springer, Cham, pp. 428-438. DOI: 10.1007/978-3-319-97909-0_46

Chollet, F., 2017. Xception: deep learning with depthwise separable convolutions. arXiv. DOI: 10.48550/arXiv.1610.02357

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L., 2009. ImageNet: a large-scale hierarchical image database. IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, pp. 248-255. DOI: 10.1109/CVPR.2009.5206848

Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A, 2010. The pascal visual object classes (VOC) challenge. International Journal of Computer Vision 88, 303–338. DOI: 10.1007/s11263-009-0275-4

Everingham, M., Eslami, S.M.A., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A, 2015. The pascal visual object classes challenge: a retrospective. International Journal of Computer Vision 111, 98-136. DOI: 10.1007/s11263-014-0733-5

Fukushima, K., 1980. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. Biological Cybernetics 36, 193–202. DOI: 10.1007/BF00344251

Girshick R., Donahue, J., Darrell, T., Malik, J., 2013. R-CNN rich feature hierarchies for accurate object detection and semantic segmentation. arXiv. DOI: 10.48550/arXiv.1311.2524

Girshick, R., 2014. Fast R-CNN. IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, pp. 1440-1448, DOI: 10.1109/ICCV.2015.169

Goldstein, A.J, Harmon, L. D., Lesk, A.B., 1971. Identification of human faces. In: Proceedings of the IEEE, vol. 59, no. 5, pp. 748-760. DOI: 10.1109/PROC.1971.8254

Guo, Y., Zhang, L., Hu, Y., He, X., Gao, J., 2016. MS-Celeb-1M: A dataset and benchmark for large-scale face recognition. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (Eds) Computer Vision – ECCV 2016. Lecture Notes in Computer Science, vol. 9907, Springer, Cham, pp 87–102. DOI: 10.1007/978-3-319-46487-9_6

He, K, Zhang, X, Ren, A., Sun, J., 2015. Deep residual learning for image recognition. arXiv. DOI: 10.48550/arXiv.1512.03385

He, K., Gkioxari, G., Dollár, P., Girshick, R., 2017. Mask R-CNN. IEEE International Conference on Computer Vision (ICCV), Venice, Italy, pp. 2980-2988. DOI: 10.1109/ICCV.2017.322

Howard, A., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T. Andreetto, A., 2017. MobileNets: efficient convolutional neural networks for mobile vision applications. arXiv. DOI: 10.48550/arXiv.1704.04861

Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K. Q., 2017. Densely connected convolutional networks. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, pp. 2261-2269. DOI: 10.1109/CVPR.2017.243

Huang, G. B., Ramesh, M., Berg, T., Learned-Miller. E., 2007. Labeled faces in the wild: a database for studying face recognition in unconstrained environments. University of Massachusetts, Amherst, Technical Report 07-49.

Jocher, G., Qiu, J., Chaurasia, A, 2023. Ultralytics YOLO (Version 8.0.0). https://github.com/ultralytics/ultralytics (Accedido 30 abril 2025).

Krizhevsky, A., Sutskever, I. Hinton, G.E., 2012. ImageNet classification with deep convolutional neural networks. Neural Information Processing Systems, 25. DOI: 10.1145/3065386.

LeCun, Y., Boser, B., Denker, J. S., Howard, R. E., Habbard, W., Jackel, L. D., Henderson, D., 1990. Handwritten digit recognition with a back-propagation network. Advances in neural information processing systems 2. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, pp. 396-404. DOI: 10.5555/109230.109279

Li, J., Wang, Y., Wan, C., Tai, Y., Qian, J., Yang, J., Wang, C., 2019. DSFD: dual shot face detector. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, pp. 5055-5064. DOI: 10.1109/CVPR.2019.00520

Lin, T. -Y., Maire, M., Belongie, S., Bourdev, L., Girshick, R., Hays, J., Perona, P., Ramanan, D., Zitnick, C. L., Dollár, P., 2015. Microsoft COCO: common objects in context. arXiv. DOI: 10.48550/arXiv.1405.0312

Nech, A., Kemelmacher-Shlizerman, I., 2017. Level playing field for million scale face recognition. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, pp. 3406-3415. DOI: 10.1109/CVPR.2017.363

Pajares, G., Herrera, P. J., Besada, E., 2021. Aprendizaje profundo. RC Libros Editorial, Madrid.

Ren S., He K., Girshick, R., Sun J., 2015. Faster R-CNN: towards real-time object detection with region proposal networks. In: Proceedings of the 29th International Conference on Neural Information Processing Systems (NIPS'15), Vol. 1. MIT Press, Cambridge, MA, USA, pp. 91–99. DOI: 10.5555/2969239.2969250

Schroff, F., Kalenichenko, D., Philbin, J., 2015. FaceNet: a unified embedding for face recognition and clustering. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, pp. 815-823. DOI: 10.1109/CVPR.2015.7298682

Simonyan, K. Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. In: 3rd International Conference on Learning Representations (ICLR 2015), San Diego, pp. 1-14. DOI: 10.48550/arXiv.1409.1556

Sirovich, L., Kirby, M., 1987. Low-dimensional procedure for the characterization of human faces. Journal of the Optical Society of America 4, 519-524. DOI: 10.1364/JOSAA.4.000519

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., 2014. Going deeper with convolutions. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, pp. 1-9. DOI: 10.1109/CVPR.2015.7298594

Tang, X., Du, D. K., He, Z, Liu, J., 2018. PyramidBox: a context-assisted single shot face detector. In: 15th European Conference on Computer Vision (ECCV 2018), Munich, Germany, Proceedings, Part IX. Springer-Verlag, Berlin, Heidelberg, pp. 812-828. DOI: 10.1007/978-3-030-01240-3_49

Turk, M., Pentland, A., 1991. Eigenfaces for recognition. Journal of Cognitive Neuroscience 3, 71-86. DOI: 10.1162/jocn.1991.3.1.71

Viola, P, Jones, M., 2001. Rapid object detection using a boosted cascade of simple features. In: Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2001), Kauai, HI, USA, pp. I-I. DOI: 10.1109/CVPR.2001.990517

Wolf, L., Hassner, T., Maoz, I., 2011. Face recognition in unconstrained videos with matched background similarity. In: IEEE Conf. on Computer Vision and Pattern Recognition (CVPR 2011), Colorado Springs, CO, USA, pp. 529-534. DOI: 10.1109/CVPR.2011.5995566

Zeiler, M., Fergus, R., 2014. Visualizing and understanding convolutional networks. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (Eds), 13th European Conference on Computer Vision (ECCV 2014), Lecture Notes in Computer Science, vol 8689, Springer, Cham. DOI: 10.1007/978-3-319-10590-1_53