Recibido 30 Abr 2026 *ReCIBE, Año 15 No. 2, junio 2026*

Aceptado 20 Jun 2026

Sistema de reconocimiento facial en video basado en Deep Learning utilizando MTCNN y VGG-16 con Transfer Learning para identificación de personajes

Deep Learning-based video face recognition system using MTCNN and VGG-16 with Transfer Learning for character identification

Rodrigo Hernandez Moncayo[1]
rhernandezm403@alumno.uaemex.mx
Alejandra Guadalupe Bravo García[1]
abravog003@alumno.uaemex.mx
Andric Javier Rodríguez Gutiérrez[1]
arodriguezg029@alumno.uaemex.mx
Antonio Ocampo Muñoz[1]
aocampom@uaemex.mx
Asdrúbal López Chau[2]
alchau@uaemex.mx

1 Centro Universitario UAEM Valle de México, México
2 Centro Universitario UAEM Zumpango, México

RESUMEN
El reconocimiento automático de rostros ha evolucionado significativamente con el uso de técnicas de aprendizaje profundo, permitiendo sistemas más robustos ante variaciones de iluminación, pose y expresión. En este trabajo se presenta la implementación de un sistema para el reconocimiento de rostros en grabaciones de video donde se utilizó el algoritmo de detección de rostros MTCNN, el cual se usa por su precisión para identificar y localizar rostros en entornos no controlados con diferentes tipos de iluminación. Tras la detección de los rostros, se empleó el modelo VGG-16 como extractor de características profundas, haciendo uso de su arquitectura para utilizar la técnica de transfer learning para aprender representaciones jerárquicas y discriminativas del rostro. El sistema procesa los cuadros del video utilizando la librería OpenCV, después se utiliza la red MTCNN para la identificación facial, posteriormente se asignaron las identidades a través de un clasificador entrenado como ya se mencionó anteriormente con el modelo VGG-16. En los estudios recientes los autores reportan que las redes convolucionales superan de forma consistente a los métodos tradicionales por su habilidad para automatizar la extracción de características relevantes y manejar condiciones complejas del entorno. Los resultados preliminares indican que la integración MTCNN–VGG-16 forma una estrategia eficaz para el reconocimiento de rostros en escenarios reales, mostrando una alta precisión y estabilidad en condiciones visuales diversas. Este trabajo aporta una arquitectura accesible y reproducible para posibles aplicaciones de análisis de video, vigilancia y automatización inteligente.

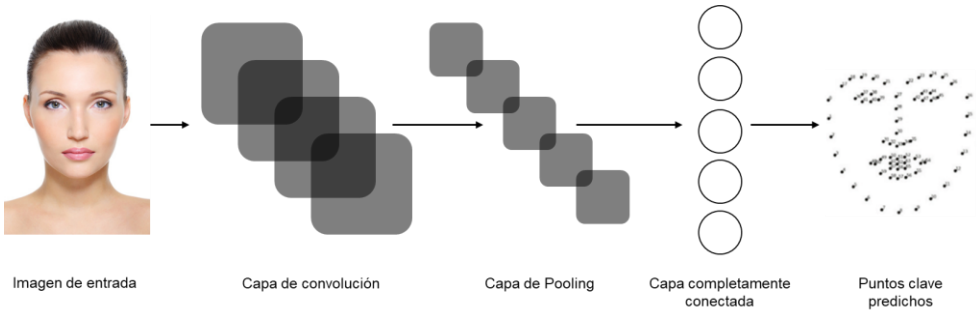
Palabras clave: Reconocimiento facial, MTCNN, VGG-16, Transfer learning, Red Neuronal Convolucional.

ABSTRACT
Automatic face recognition has evolved significantly with the use of deep learning techniques, enabling more robust systems that can withstand variations in lighting, pose, and expression. This paper presents the implementation of a face recognition system for video recordings using the MTCNN face detection algorithm, which is used for its accuracy in identifying and locating faces in uncontrolled environments with varying lighting conditions. After face detection, the VGG-16 model was used as a deep feature extractor, leveraging its architecture to employ transfer learning techniques to learn hierarchical and discriminative representations of the face. The system processes the video frames using the OpenCV library, then uses the MTCNN network for facial identification, and finally assigns identities through a classifier trained, as previously mentioned, with the VGG-16 model. In recent studies, authors report that convolutional neural networks consistently outperform traditional methods due to their ability to automate the extraction of relevant features and handle complex environmental conditions. Preliminary results indicate that the MTCNN–VGG-16 integration forms an effective strategy for facial recognition in real-world scenarios, demonstrating high accuracy and stability under diverse visual conditions. This work provides an accessible and reproducible architecture for potential applications in video analytics, surveillance, and intelligent automation.

Keywords: Face recognition, MTCNN, VGG-16, Transfer learning, Convolutional Neural Networks

1. INTRODUCCIÓN

El reconocimiento facial es una de las áreas más relevantes dentro de la visión por computadora, impulsado por su utilidad en seguridad, autenticación biométrica, interacción humano-computadora y análisis de contenido audiovisual (Kortli et al., 2020). Los métodos tradicionales, tales como Eigenfaces (técnica que representa los rostros mediante vectores propios obtenidos a partir del análisis estadístico de imágenes faciales), LBP (Local Binary Patterns, Patrones Binarios Locales utilizados para describir la textura de la imagen facial) o PCA (Principal Component Analysis, Análisis de Componentes Principales, empleado para reducir la dimensionalidad de los datos y resaltar las características más significativas), fueron durante años la base de los sistemas de reconocimiento facial (Wang, 2024); sin embargo, su desempeño se degrada significativamente en entornos no controlados debido a variaciones en iluminación, pose, expresiones faciales y oclusiones (Southwest Minzu University, Key Laboratory of Electronic and Information Engineering, State Ethnic Affairs Commission, Chengdu, 610041, China et al., 2020). Las redes neuronales profundas, y particularmente las redes convolucionales (CNN), permiten aprender automáticamente características representativas sin necesidad de diseñarlas manualmente, esto a través de múltiples capas jerárquicas, en las cuales se extrae características básicas como bordes, contornos y texturas; posteriormente, capas más profundas identifican patrones complejos relacionados con rasgos faciales como ojos, nariz y boca; y finalmente, estas características se transforman en una representación vectorial (embedding) que permite la identificación y clasificación precisa del rostro, como lo representa la **Figura 1**. Esto ha generado un salto significativo en precisión y robustez para el reconocimiento facial (Li et al., 2020). Modelos basados en CNN han demostrado ser capaces de manejar variaciones de pose, iluminación, distancia y resolución, convirtiendo a las CNN en el método estándar para la extracción de características faciales (Fuad et al., 2021).



Fuente: Elaboración propia.

Figura 1 Representación del reconocimiento facial mediante una CNN.

En este marco, la detección facial constituye un componente crítico. El algoritmo MTCNN ha demostrado ser uno de los detectores más eficientes, dado que realiza detección y alineación simultáneamente mediante tres redes en cascada (P-Net, R-Net y O-Net), alcanzando alta precisión y funcionamiento en tiempo real (Zhang et al., 2016). Para la identificación posterior, modelos como VGGNet/VGG-16 son ampliamente utilizados gracias a su arquitectura profunda que emplea filtros convolucionales pequeños (3×3) y múltiples capas para capturar

representaciones jerárquicas del rostro (Parkhi et al., 2015). Estos modelos son particularmente aptos para estrategias de *transfer learning*, permitiendo adaptar redes preentrenadas a nuevos individuos o conjuntos específicos sin requerir millones de imágenes (Southwest Minzu University, Key Laboratory of Electronic and Information Engineering, State Ethnic Affairs Commission, Chengdu, 610041, China et al., 2020).

El presente trabajo tiene como objetivo desarrollar e implementar un sistema de reconocimiento facial en video basado en técnicas de aprendizaje profundo que combine el algoritmo MTCNN para la detección y alineación de rostros con la arquitectura VGG-16 como extractor de características, con el fin de identificar personas de manera precisa y robusta en entornos no controlados.

2. MARCO TEÓRICO

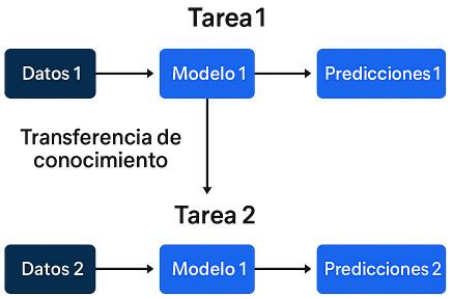
El reconocimiento automático de rostros es una de las áreas más consolidadas dentro de la visión por computadora debido a su importancia en aplicaciones de vigilancia, autenticación, análisis de comportamiento y sistemas inteligentes. El reconocimiento facial es una parte importante de los sistemas de visión por computadora y que es derivado del avance de la inteligencia artificial. Estas tecnologías han habilitado la automatización de procesos de identificación mediante la extracción de patrones biométricos, aumentando su uso en diversos sectores de la tecnología (Sanabria Moyano et al., 2022). En las últimas décadas este campo ha evolucionado, habiéndose utilizado desde enfoques estadísticos clásicos hasta arquitecturas profundas y altamente optimizadas, que permiten una mayor robustez ante variaciones en el entorno, como iluminación, pose, oclusiones y condiciones y defectos reales de captura.

En el pasado del reconocimiento facial, el panorama estaba dominado por métodos basados en aproximaciones estadísticas como Eigenfaces, que utilizaban el algoritmo de análisis de componentes principales PCA para descomponer en componentes una imagen y reducir su dimensionalidad, y técnicas como Elastic Graph Matching, que representaban el rostro como una estructura deformable (Wang, 2024). Sin embargo, estos enfoques en escenarios no controlados presentaban limitaciones importantes, pues la calidad de las características extraídas se veía degradada por la variabilidad del entorno. (Southwest Minzu University, Key Laboratory of Electronic and Information Engineering, State Ethnic Affairs Commission, Chengdu, 610041, China et al., 2020) Posteriormente, la comunidad científica adoptó descriptores locales como Local Binary Patterns (LBP) y Gabor Filters para la extracción de patrones texturales, que aumentaron la robustez del sistema, aunque sin lograr generalizar de manera óptima frente a grandes variaciones inter e intra sujeto (Fuad et al., 2021).

Por otra parte, el aprendizaje profundo (deep learning) es una rama del aprendizaje automático basada en redes neuronales artificiales con múltiples capas, capaces de aprender representaciones jerárquicas a partir de datos como imágenes o video (Sanchez & Sigua, s. f.). Entre los algoritmos que existen dentro del Deep learning están las Redes Neuronales Convolucionales (CNN), las cuales fueron una revolución moderna del reconocimiento facial, cuyo principio fundamental es la extracción jerárquica y automática de características relevantes directamente desde los píxeles de la imagen, sin necesidad de ingeniería manual de atributos (Krichen, 2023). Las CNN demostraron superar significativamente a los métodos clásicos en tareas de clasificación de imágenes, detección de objetos y reconocimiento facial, dado que son capaces de aprender representaciones invariantes a transformaciones locales (Gupta et al., 2022). Estas arquitecturas profundas evolucionaron hacia modelos de gran escala como VGG-16, FaceNet, ResNet y

derivados, que alcanzaron niveles de precisión superiores al rendimiento humano en algunos benchmarks especializados.

Un componente clave del éxito de las CNN radica en la disponibilidad de grandes bases de datos etiquetadas que permiten ajustar millones de parámetros (Krizhevsky et al., 2012). No obstante, en áreas específicas, como el reconocimiento de personas en videos particulares, esas bases de datos no siempre existen o resultan costosas de construir. Este problema motivó el desarrollo de Transfer Learning, una estrategia que reutiliza modelos previamente entrenados en grandes datasets para acelerar el aprendizaje en tareas nuevas con datos limitados (Hosna et al., 2022), como se representa en la **Figura 2**, donde en la Tarea 1, el modelo es entrenado con un primer conjunto de datos (Datos 1), permitiéndole aprender representaciones generales útiles para la resolución del problema, generando las Predicciones 1; posteriormente, dicho conocimiento se transfiere hacia la Tarea 2, donde el mismo modelo preentrenado se reutiliza y se adapta a un nuevo conjunto de datos (Datos 2); en esta segunda fase, no se entrena el modelo desde cero, sino que se aprovechan sus parámetros previamente aprendidos, lo que permite reducir el tiempo de entrenamiento, mejorar la precisión y favorecer una mejor generalización en tareas similares. Conceptualmente, el aprendizaje por transferencia explota el conocimiento adquirido en un "dominio fuente" para mejorar el desempeño en un "dominio objetivo", reduciendo los costos de entrenamiento y mejorando la generalización (Hosna et al., 2022).



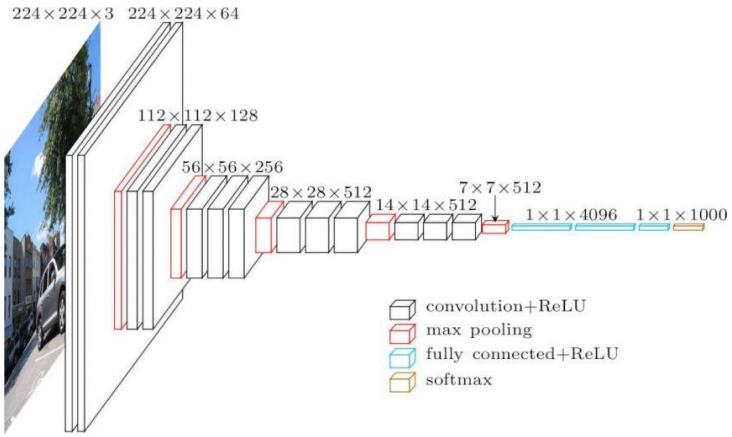
Fuente: Elaboración propia.

Figura 2 Representación gráfica del Transfer Learning entre tareas.

También se han propuesto arquitecturas que integran redes CNN con modelos secuenciales capaces de capturar dependencias temporales. En particular, el uso de VGG-16 en un esquema Time-Distributed permite procesar cada fotograma de forma uniforme, extrayendo características espaciales consistentes que posteriormente son analizadas por redes LSTM, las cuales modelan la secuencia completa al aprender relaciones contextuales entre cuadros consecutivos (Athira & Divya Udayan, 2024), (Orozco et al., 2020), (Srabu & Santra, 2021).

Asimismo, investigaciones recientes han evidenciado que el uso de VGG-16 como extractor de características, combinado con técnicas de transferencia de aprendizaje, no solo mejora la discriminación entre identidades faciales, sino que también incrementa la eficiencia computacional y la capacidad de operación en tiempo real (Paulchamy et al., 2025), (Dar & Palanivel, 2021). En sistemas de detección facial, VGG-16 ha demostrado un elevado desempeño incluso bajo condiciones no controladas, como rostros parcialmente cubiertos, variaciones de pose y fondos complejos. Gracias a su arquitectura profunda basada en filtros convolucionales jerárquicos, que se representa en la **Figura 3**, este modelo logra captar patrones faciales complejos

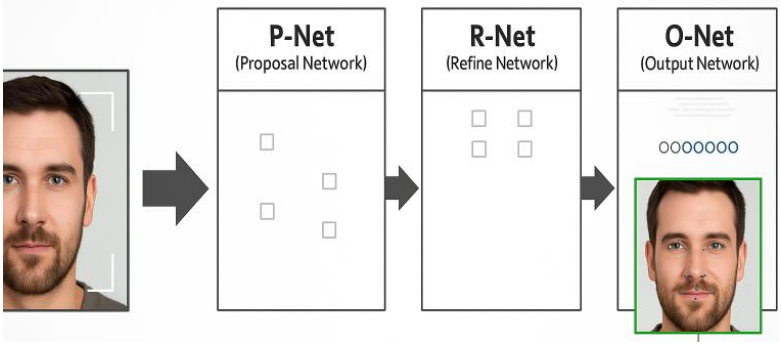
desde niveles bajos hasta estructuras de alto nivel, generando representaciones robustas que facilitan la identificación precisa (Vimal & Shirivastava, 2022).



Fuente: Hassan, M. U. (2023, 9 mayo). VGG16 – Convolutional Network for Classification and Detection. Neurohive / Neural Networks. https://neurohive.io/en/popular-networks/vgg16/

Figura 3 Arquitectura VGG-16.

Otra área fundamental dentro del pipeline es la detección de rostros, cuya finalidad es localizar y recortar la región facial dentro de una escena. Entre las técnicas más influyentes se encuentra MTCNN (Multi-Task Cascaded Convolutional Neural Network), en la **Figura 4**, propuesta como un sistema de cascada que combina tres redes convolucionales: P-Net, R-Net y O-Net. Estas redes operan de forma jerárquica para realizar simultáneamente la detección y alineación facial a través de regresión de bounding boxes y estimación de landmarks (Southwest Minzu University, Key Laboratory of Electronic and Information Engineering, State Ethnic Affairs Commission, Chengdu, 610041, China et al., 2020). MTCNN destaca por su capacidad para operar en tiempo real, su bajo consumo de memoria y su robustez ante variaciones de iluminación, posición y expresiones faciales. Gracias a su diseño multitarea, MTCNN se ha convertido en un estándar para la etapa inicial del proceso de reconocimiento facial en sistemas basados en deep learning (Zhang et al., 2016), (Xie et al., 2020).



Fuente: Elaboración propia.

Figura 4 Técnica MTCNN.

Además de la detección, otro aspecto relevante es la extracción de características profundas, particularmente a través de modelos preentrenados diseñados para tareas de identificación facial (Almabdy & Elrefaei, 2019), (Hassanat et al., 2021). Uno de los modelos más influyentes es VGG-16, el cual emplea una arquitectura de redes convolucionales profundas inspirada en VGG-16 y entrenada en millones de imágenes de rostros humanos. Este modelo ha demostrado producir representaciones vectoriales faciales sumamente discriminativas, con un alto rendimiento en bases como LFW, CASIA-WebFace y otras (Ardiawan & Negarara, 2024). Otras investigaciones han demostrado que las redes preentrenadas para el reconocimiento facial pueden ser reutilizadas exitosamente aplicándose a tareas como la estimación de edad y la detección de emociones (Qawaqneh et al., 2017), lo cual pone de manifiesto la capacidad que tienen para aprender representaciones generales y útiles para diferentes fines.

La literatura reciente también ha mostrado un crecimiento significativo en la integración de técnicas de deep learning con redes especializadas, funciones de pérdida avanzadas y estrategias de regularización para mejorar la separación entre clases similares y la estabilidad del modelo (Zhong et al., 2021). A su vez, el uso de arquitecturas residuales profundas, métodos de data augmentation y técnicas de alineación facial han fortalecido la robustez del reconocimiento frente a ruido, desenfoque y oclusiones parciales (Cao et al., 2018), (Ding et al., 2017). En un contexto moderno de las aplicaciones para el reconocimiento facial en video, diversos desafíos adicionales son presentados en relación a que es necesaria una detección continua frame por frame y un manejo de variaciones temporales además del procesamiento eficiente en tiempo real (Nguyen, et al., 2017).

Estudios recientes resaltan la importancia de la extracción de características consistentes mediante múltiples fotogramas, también destacan el tracking facial y el uso de pipelines optimizados para flujos multimedia (Fuad et al., 2021). En este tipo de escenarios, la integración de MTCNN para la detección y alineación de los rostros y VGG-16 para la extracción de representaciones vectoriales, forman una solución robusta, en donde a partir de estos elementos es factible construir sistemas que sean capaces de reconocer individuos específicos en secuencias de videograbaciones, manteniendo un desempeño confiable a pesar de las condiciones variables del entorno y la disponibilidad limitada de datos

3. ARQUITECTURA DEL SISTEMA Y METODOLOGÍA

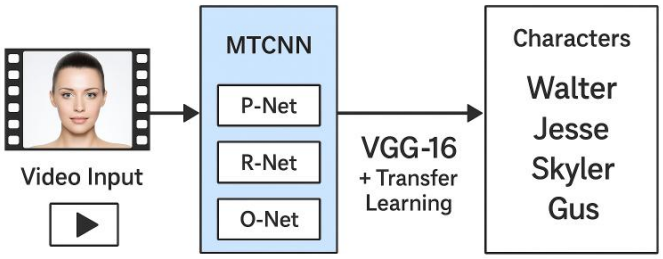
La arquitectura propuesta para el sistema de reconocimiento facial se estructura como un pipeline modular compuesto por tres etapas principales: adquisición del dataset, extracción profunda de características y clasificación de la identidad. Cada uno de estos módulos se implementó utilizando herramientas de aprendizaje profundo, visión por computadora y modelos preentrenados, permitiendo un funcionamiento robusto en videos reales. En este artículo, se toma como base la serie Breaking Bad (Gilligan, 2008) para realizar el reconocimiento facial de los personajes Walter, Jesse, Gus, Skyler.

3.1. Arquitectura propuesta

La **Figura 5** muestra la arquitectura final del sistema propuesto para el reconocimiento facial en video. El proceso inicia con la entrada de un video, del cual se extraen frames de manera secuencial. Cada frame es procesado por el algoritmo MTCNN, encargado de detectar y localizar automáticamente los rostros presentes. MTCNN consiste de las siguientes tres etapas en cascada:

a) P-Net, que genera regiones candidatas; b) R-Net, que refina las detecciones eliminando falsos positivos; y c) O-Net, que realiza la detección final junto con la alineación facial.

Posteriormente, los rostros detectados son redimensionados y son enviados al modelo preentrenado VGG-16, al cual se le aplicó aprendizaje por transferencia para adaptar sus últimas capas para la identificación de personajes de la serie de televisión Breaking Bad. De esta forma, el sistema genera como salida la clase a la cual corresponde el personaje identificado junto con su porcentaje de confianza, habilitando así la asociación automática en tiempo real dentro del flujo del video para cada personaje.



Fuente: Elaboración propia.

Figura 5 Arquitectura final del sistema de reconocimiento facial en video basado en MTCNN y VGG-16 con Transfer Learning.

3.2. Selección del dataset

Para la implementación de la red de reconocimiento facial de la serie de televisión Breaking Bad, se utilizó un data set con 5 clases que son: Walter, Jesse, Guss, Skyler y Desconocido. Por cada clase se tienen 3000 imágenes, dando un total de 15000 imágenes a color de 256 x 256 pixeles. El dataset se dividió en 80% para entrenamiento, 10% para validación y 10% para pruebas. Para obtener el data set se utilizaron vídeos de la plataforma YouTube con una resolución de 1920 x 1080p (Full High Definition) a 30 fps (frame x secod), con diferentes escenas de la serie. Además, solo se fueron guardando uno de cada 10 fotogramas para obtener imágenes más variadas. Para el procesamiento del dataset se utilizó la plataforma de Google Colab con una cuenta gratuita, dónde se utilizó la red MTCNN. Según sus autores esta fue entrenada con los datasets WinterFace y CelebA, teniendo un total de 800000 imágenes de diferentes caras, ángulos e iluminación.

Los videos se procesaron obteniendo los fotogramas y al mismo tiempo se implementaba MTCNN, esta red se especializa en detectar por medio de puntos las diferentes características del rostro como lo son los ojos derecho e izquierda, la nariz, las esquinas de la boca izquierda y derecha, lo valioso de esta red es que puede detectar múltiples rostros en una sola escena, con diferentes tipos de iluminación y ángulos.

La tabla 1 muestra un resumen del conjunto de datos generado.

Característica	Valor
Total de muestras	15000
Categorías	Walter, Jesse, Guss, Skyler y Desconocido
Tamaño de los frames	256 x 256 pixeles
Modelo de color	RGB

Tabla 1. Conjunto de datos generado para entrenamiento y prueba del modelo

La red está compuesta por tres redes convolucionales que están conectadas en cascada las cuales funcionan de la siguiente manera:

P-Net (Proposal Network): Esta red mapea las posibles regiones de la imagen donde se puede encontrar uno o varios rostros, y aunque no lo hace de manera precisa devuelve los bounding boxes (cajas delimitadoras) con las diferentes posibilidades.

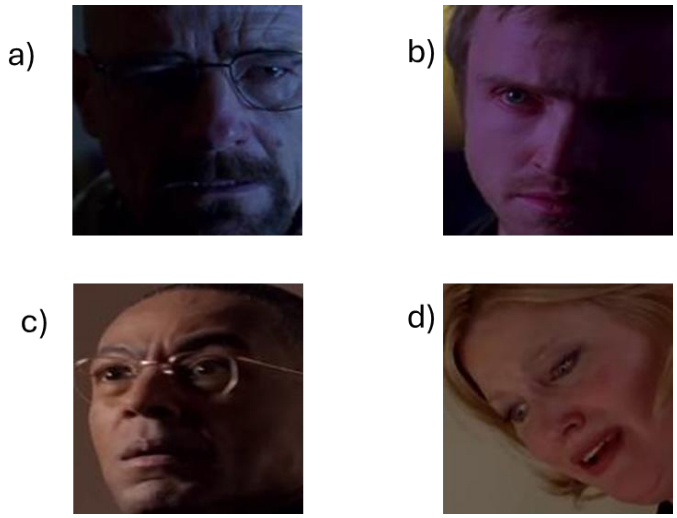
R-Net (Refinement Network): refina la búsqueda de los rostros encontrados por P-Net, esto quiere decir que detecta falsos positivos descartando imágenes dónde no se encuentran caras.

O-Net (Output Network): en esta red se realiza la predicción de los puntos de las características principales del rostro, se calculan las coordenadas y se realiza el último ajuste de las bounding boxes.

En la **Figura 6**, se presenta el ejemplo de las imágenes obtenidas para el entrenamiento de la red. Como se puede observar en los rostros se pudieron obtener aun en diferentes tipos de iluminación y ángulos.

En las imágenes del inciso a y b se puede ver el rostro de Walter y Jesse de frente y la iluminación no es del todo clara y los rostros solo se pueden observar de manera parcial. En el inciso c (Gus) y d (Skyler) se ven los rostros bien iluminados, aunque en diferente ángulo. En el caso de Gus el rostro se encuentra girado tres cuartos y en el caso de Skyler se encuentra a 45 grados, pero mirando hacia abajo.

Esto demuestra que la red MTCNN se comporta de una manera robusta y es capaz de detectar los rostros aun en condiciones que no son óptimas, sin embargo, esta característica hace que aun rostros que no están bien definidos sean detectados. Las imágenes que no logran identificarse bien son desechadas manualmente ya que no sirven para el propósito de esta investigación.



Fuente: Elaboración propia.

Figura 6 Ejemplo de las imágenes utilizadas en el dataset. Se muestran los rostros que se ocuparon para el entrenamiento y detección en los videos, obtenidos de la serie Breaking Bad de los siguientes personajes: a) Walter, b) Jesse, c) Gus y d) Skyler.

3.3. Entrenamiento Transfer Learning

Una vez que se ha creado la base de datos, el sistema procede a la extracción de características de los rostros utilizando un modelo VGG pre-entrenado ajustado mediante transfer learning. Este modelo derivado de la arquitectura VGG-16 posee 16 capas convolucionales que contiene 138 millones de parámetros. Y aunque es una red bastante robusta y uniforme, también es cierto que por su simplicidad en cuanto a los filtros (1 x 1, 3 x 3) se continúa usando actualmente. Existen en la arquitectura varias capas de pooling que hace que la profundidad de la red se reduzca. En cuanto al número de filtros se pueden usar 64, que se pueden duplicar en 128 y luego a 256. En las últimas capas, podemos usar 512 filtros.

Esta red acepta como entrada imágenes de 224 x 224 pixeles de 3 canales RGB. En este trabajo se utilizaron imágenes de 256 x 256, que se redimensionan en el preprocesamiento y normalización. Para utilizar la técnica de transfer learning las 16 capas se mantienen congeladas durante el entrenamiento base, para aprovechar las características de bajo nivel que ya ha aprendido la red previamente. Al evitar un entrenamiento exhaustivo permite reducir el tiempo de entrenamiento y mejorar la capacidad de generalizar del nuevo modelo ya que se reutilizan los pesos pre-entrenados, este método es usado en situaciones cuando se dispone de un conjunto limitado de imágenes como es este caso.

El proceso de congelar las capas permitió que el modelo se especializara en la clasificación de cinco clases, se agregan nuevas capas densas específicas para la clasificación para evitar sobre ajuste y mejorar la eficiencia:

1. Una capa Flatten convierte la salida de las capas convolucionales en un vector unidimensional, necesario para las capas densas.
2. Una capa densa de 256 neuronas con activación ReLU introduce no linealidad y permite que el modelo aprenda las características del nuevo dataset.

3. Se aplica una capa de Dropout con una tasa de 0.5 y aumento de datos (Image Data Generador) para evitar el sobreajuste, esto mejora la capacidad de generalizar el aprendizaje del modelo propuesto.
4. En la parte final se usa una capa Densa con 5 neuronas y activación softmax para realizar la clasificación de las 5 clases utilizadas.
5. La red fue entrenada por 30 épocas donde se guardó el mejor modelo basado en la pérdida. Para después proceder al entrenamiento por medio de fine-tuning (ajuste fino) por 20 épocas más,

Sin embargo, el modelo aún podía mejorar en términos de rendimiento específico con el fin de ajustarse a las características y peculiaridades del dataset. Por lo tanto, se hizo un proceso de fine-tuning en las capas superiores del modelo. De esta forma, las dos capas finales de las 16 con las que cuenta el VGG16 fueron descongeladas para permitir que sus pesos fueran actualizados durante el entrenamiento.

El fine-tuning fue realizado con una tasa de aprendizaje significativamente más baja (1e-5) en comparación con el entrenamiento inicial, utilizando el optimizador RMSprop. Esto facilita una mejor generalización del modelo. Al utilizar una tasa de aprendizaje más baja se asegura que en las capas previamente descongeladas se realicen ajustes pequeños los cuales evitan que se modifique el conocimiento que ya adquirió la red en el preentrenamiento.

Durante esta fase, se implementaron dos técnicas para controlar el entrenamiento y evitar el sobreajuste de la red. La primera fue *EarlyStopping*, esta permite detener el entrenamiento si no mejora la pérdida de validación *(val_loss)* después de 5 épocas. La segunda técnica que se utilizó fue la de *ModelCheckpoint*, que guarda los pesos del modelo con el mejor rendimiento en el conjunto de validación. Esto asegura que el modelo más eficiente sea guardado en formato h5.

La tabla 2 muestra un resumen de los principales hiper-parámetros usados en el modelo.

Hiper-parámetro	Valor
Tamaño de la entrada	256 x 256 x 3 redimensionado internamente a 224 x 224 x 3
Optimizador	Adam
Learning rate	1e-3 para entrenamiento inicial 1e-5 para fine tunning
Capas	Las predeterminadas para VGG-16
Batch size	20
Épocas	30 para fine tuning

Tabla 2. Hiper-parámetros utilizados

Por las ventajas mencionadas a lo largo de este artículo, la combinación de MTCNN y VGG-16 mejora la precisión de la red implementada, debido a que la primera red garantiza detecciones alineadas y confiables, mientras que la segunda red genera representaciones profundas que logran

separar de manera eficiente identidades faciales incluso en condiciones visuales adversas. La arquitectura modular permite además ampliar el sistema hacia nuevas identidades o modelos alternativos, lo que la convierte en una solución adaptativa para contextos reales de análisis automático de rostros en video.

4. RESULTADOS

En esta sección se presentan los resultados del entrenamiento por medio de transfer learning, como se puede observar en la **Figura 7**, donde se representa del lado izquierda la precisión y del lado derecho la pérdida. En ambas se muestran una curva azul que representan el comportamiento durante el entrenamiento y una curva naranja que muestra el desempeño en el data set de validación.

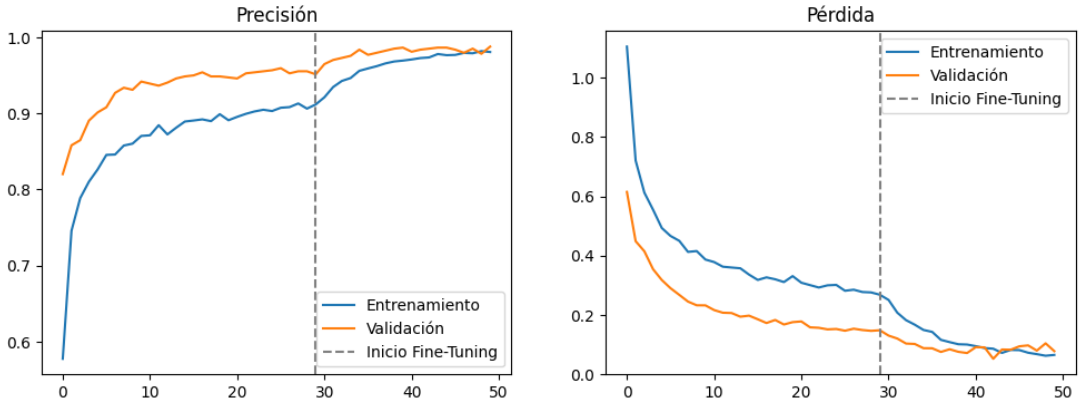
Como se mencionó anteriormente, el entrenamiento de la red se realizó por 50 épocas (eje X) en dos etapas, 30 épocas iniciales para el entrenamiento base y las siguientes 20 épocas por medio de la técnica fine tuning delimitadas por la línea negra puntada.

En la gráfica de precisión se puede observar como la curva azul empieza en 0.3 y va subiendo gradualmente la exactitud del modelo conforme va aprendiendo a clasificar de forma correcta los datos en el conjunto de entrenamiento, por su parte, la curva naranja que representa el conjunto de validación, comienza más alta (aproximadamente 0.8) y también va subiendo lentamente.

En la gráfica de pérdida, la curva azul inicia su trayecto arriba de 1.0 y va disminuyendo rápidamente, esto indica que el modelo va mejorando y se está ajustando de forma correcta a los datos de entrenamiento. Mientras que la curva naranja empieza alrededor de 0.6 y disminuye lentamente hasta terminar en 0.1. Esto indica que el modelo está obteniendo la capacidad de generalización.

Antes de la época 30, el modelo entrena con una tasa de mejora en precisión y reducción de pérdida, tanto en entrenamiento como en validación. Después de la época 30 (punto de fine-tuning), la precisión de validación mejora más lentamente y la pérdida se estabiliza. El fine-tuning ajusta el modelo para mejorar estas métricas.

Al referenciar los valores de las gráficas junto con los puntos de las curvas, podemos ver cómo el modelo se va ajustando a lo largo del tiempo y cómo el fine-tuning impacta en las métricas clave.

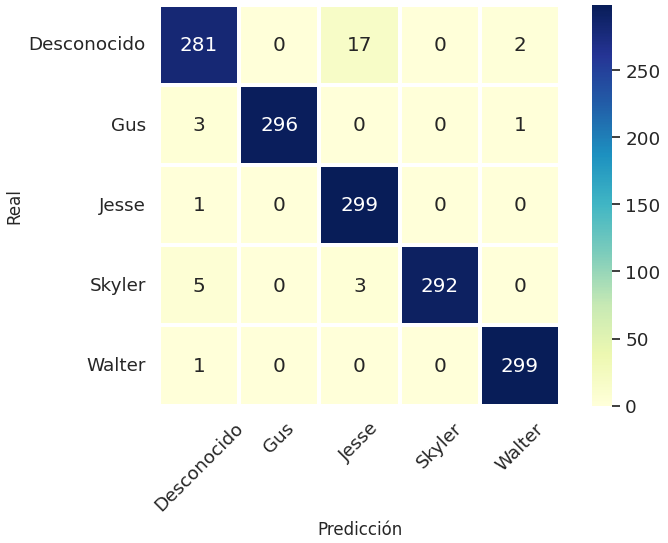


Fuente: Elaboración propia.

Figura 7 Curvas de precisión y pérdida antes y después del Fine-Tuning.

La matriz de confusión es una herramienta que se utilizo para evaluar el comportamiento del modelo implementado, las celdas muestran el numero de instancias que fueron clasificadas en una clase particular. La matriz presentada en la **Figura 8**, proporciona una evaluación visual del desempeño del modelo de clasificación de las cinco categorías utilizadas: *Desconocido, Gus, Jesse, Skyler y Walter*, utilizando imágenes (data_test) que no han sido utilizadas para el entrenamiento, ni la validación.

En la diagonal de color azul rey se pueden observar las predicciones que son correctas del modelo, donde se puede observar que el modelo tiene un nivel alto de exactitud, al clasificar las imágenes que no había visto anteriormente. Sin embargo, fuera de la diagonal, se pueden observar ciertos errores de clasificación que ofrecen información sobre áreas de mejora.



Fuente: Elaboración propia.

Figura 8 Matriz de confusión del desempeño del modelo.

Es importante notar que los errores de clasificación son mínimos, lo que sugiere que el modelo está generalizando bien la información. Sin embargo, existen errores entre las categorías Desconocido y Jesse, así como de manera más baja en Skyler y Desconocido. Esto podría indicar que entre estas clases al modelo le es difícil diferenciarlas.

La **Figura 9**, presenta una serie de ejemplos representativos de las predicciones generadas por el sistema sobre imágenes individuales extraídas del conjunto de prueba. En cada recuadro se indica la clase real y la clase predicha, acompañadas del porcentaje de confianza correspondiente. Se observa que el modelo logra identificar correctamente al personaje Jesse en múltiples condiciones de iluminación y expresiones faciales; sin embargo, también se evidencian algunos casos donde rostros pertenecientes a la clase Desconocido son erróneamente clasificados como Jesse, lo que confirma la existencia de confusión entre estas clases, tal como se reflejó previamente en la matriz de confusión.

Fuente: Elaboración propia.

Figura 9 Ejemplos de predicciones del modelo sobre imágenes del conjunto de prueba.

Observando la matriz de confusión se puede concluir que el modelo tiene un desempeño sobresaliente al realizar las predicciones correctas en todas las clases, lo cual infiere que se tiene una alta generalización del modelo.

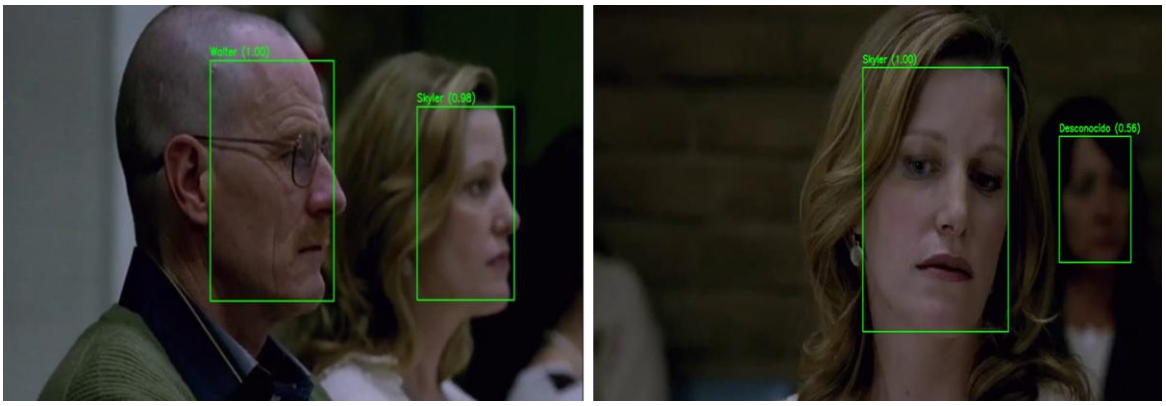
Para la prueba de clasificación se tomó un video en FHD de la plataforma Youtube y se editó para que tuviera la duración no mayor a un minuto, eliminando el canal de audio para evitar que el peso del video no supere los 200Mb. Esto para que Colab pueda utilizar la red MTCC y las predicciones al mismo tiempo.

La entrada para la predicción se usó un archivo de video en formato mp4, para el procesamiento del video se utiliza la librería OpenCV, en esta etapa se obtiene todos las frames del video para en seguida convertirlos a formato RGB, redimensionando la imagen a 256 x 256. Posteriormente, la imagen es normalizada y convertida en un array para ser procesada por el modelo. Como ya se mencionó en la etapa de entrenamiento para la detección de los rostros se utilizó la red MTCNN. Esta etapa es crucial, ya que garantiza la localización de múltiples rostros en un frame antes de aplicar la clasificación.

Una vez que los rostros han sido detectados en el video, se procede a cargar el archivo h5 con el modelo entrenado y se ajusta para predecir la identidad de los rostros detectados de acuerdo con las 5 clases con las que entrenó (Desconocido, Gus, Jesse, Skyler y Walter). La salida del modelo es un vector de probabilidades, el cual fue utilizado para determinar la identidad del rostro de acuerdo al mayor número entregado.

De esta forma, el resultado de la clasificación es impreso en el frame original, mostrando recuadros de color verde que enmarcan cada rostro detectado, el nombre de la clase perteneciente y el porcentaje asociado a la predicción. Este proceso es repetido para cada 50 frames del video. Finalmente, el video que contiene las predicciones es guardado en formato mp4 para la revisión posterior de los resultados. El resultado se pude observar en la **Figura 10**, donde en la escena del lado izquierdo se muestra al personaje Walter con una predicción del 100%, mientras que en la escena derecha se observa a Skyler junto a un rostro clasificado como Desconocido, donde se

muestra cómo el sistema es capaz de detectar múltiples rostros simultáneamente y asignar un valor de probabilidad a cada clase incluso bajo condiciones de iluminación variables.



Fuente: Elaboración propia.

Figura 10 Resultado del reconocimiento facial en video con bounding boxes y predicción de identidad.

5. DISCUSIÓN

La implementación propuesta demostró que la combinación de MTCNN para la detección y VGG-16 con transferencia de aprendizaje para la extracción de características ofrece un desempeño sólido en el reconocimiento facial en video. Los resultados de precisión y pérdida muestran un comportamiento estable durante el entrenamiento, lo que coincide con lo reportado en la literatura, donde las CNN y el transfer learning permiten mejorar la generalización incluso con conjuntos de datos limitados. En este sentido se puede observar mediante la disminución paulatina de las métricas de pérdida y el aumento gradual de la precisión, que el modelo fue capaz de aprender representaciones profundamente discriminativas para cada una de las cinco clases utilizadas.

La matriz de confusión es la prueba de esta tendencia, encontrándose distribuidas la mayoría de las predicciones en la diagonal principal. Sin embargo, se observan algunos errores de predicción entre las clases Desconocido y Jesse, así como confusiones menores con Skyler. Esto se atribuye a la similitud visual en ciertas escenas, la variación en la iluminación y ángulos que afectan la calidad de las imágenes en el conjunto de datos. Esto coincide con estudios previos en donde se señala que, aun con modelos profundos, la clasificación de rostros con características similares sigue siendo un desafío especialmente en secuencias de video.

Por otra parte, la estabilidad del modelo posterior al fine-tuning sugiere que la estrategia de descongelar solamente las capas superiores fue adecuada para evitar sobreajustes y mejorar el rendimiento sin comprometer el conocimiento que ya se tiene del preentrenamiento. En esta línea las pruebas hechas en video demostraron que el sistema es capaz de mantener un desempeño confiable en escenarios dinámicos donde existen cambios temporales, múltiples rostros y condiciones visuales cambiantes

De forma general, los resultados obtenidos corroboran que el enfoque propuesto es apropiado para tareas de reconocimiento facial en video y que se pueden considerar como una base sólida

para mejoras en el trabajo a futuro, como la ampliación del dataset y la integración de técnicas de traking temporal.

6. CONCLUSIONES

El sistema desarrollado en esta investigación demuestra que la integración de MTCNN y VGG-16 es adecuada para tareas de reconocimiento facial en video donde no se tiene un control total de entorno, ya que permitió procesar secuencias de frames en condiciones reales con un alto nivel de estabilidad. De la misma forma, la arquitectura propuesta evidenció que es posible obtener modelos precisos aun trabajando con un conjunto de datos moderado, siempre y cuando se empleen las técnicas adecuadas de transferencia de aprendizaje y estrategias de regularización. En este caso, los errores identificados entre ciertas clases proponen que la variabilidad del dataset sigue siendo un factor de impacto en los resultados, por lo que ampliar el número de muestras y diversificar las escenas podría fortalecer la robustez del sistema. En conjunto, los resultados respaldan la factibilidad del enfoque para aplicaciones prácticas y generan las bases para futuras mejoras en escenarios con mayor complejidad y diversidad, fortaleciendo su aplicabilidad en contextos reales cada vez más exigentes.

Referencias

Almabdy, S., & Elrefaei, L. (2019). Deep Convolutional Neural Network-Based Approaches for Face Recognition. Applied Sciences, 9(20), 4397. https://doi.org/10.3390/app9204397

Ardiawan, M. I., & Negarara, G. P. K. (2024). A Comparative Analysis of FaceNet, VGGFace, and GhostFaceNets Face Recognition Algorithms For Potential Criminal Suspect Identification. Journal of Applied Artificial Intelligence, 5(2), 34-49. https://doi.org/10.48185/jaai.v5i2.1237

Athira, K. A., & Divya Udayan, J. (2024). Temporal Fusion of Time-Distributed VGG-16 and LSTM for Precise Action Recognition in Video Sequences. Procedia Computer Science, 233, 892-901. https://doi.org/10.1016/j.procs.2024.03.278

Cao, K., Rong, Y., Li, C., Tang, X., & Loy, C. C. (2018). Pose-robust face recognition via deep residual equivariant mapping. arXiv. https://arxiv.org/abs/1803.00839

Dar, S. A., & Palanivel, S. (2021). Performance evaluation of convolutional neural networks (CNNs) and VGG on real time face recognition system. Advances in Science, Technology and Engineering Systems Journal, 6(2), 956–964. https://doi.org/10.25046/aj0602109

Ding, Y., Cheng, Y., Cheng, X., Li, B., You, X., & Yuan, X. (2017). Noise-resistant network: a deep-learning method for face recognition under noise. EURASIP Journal on Image and Video Processing, 2017, Article 43. https://doi.org/10.1186/s13640-017-0188-z

Fuad, Md. T. H., Fime, A. A., Sikder, D., Iftee, Md. A. R., Rabbi, J., Al-Rakhami, M. S., Gumaei, A., Sen, O., Fuad, M., & Islam, Md. N. (2021). Recent Advances in Deep Learning Techniques for Face Recognition. IEEE Access, 9, 99112-99142. https://doi.org/10.1109/ACCESS.2021.3096136

Gilligan, V. (Creador). (2008–2013). *Breaking Bad* [Serie de televisión]. High Bridge Productions; Gran Via Productions; Sony Pictures Television.

Gupta, J., Pathak, S., & Kumar, G. (2022). Deep Learning (CNN) and Transfer Learning: A Review. Journal of Physics: Conference Series, 2273(1), 012029. https://doi.org/10.1088/1742-6596/2273/1/012029

Hassanat, A. B., Albustanji, A., Tarawneh, A. S., Alrashidi, M., Alharbi, H., Alanazi, M., Alghamdi, M., Alkhazi, I. S., & Prasath, V. B. S. (2021). Deep learning for identification and face, gender, expression recognition under constraints. arXiv. https://arxiv.org/abs/2111.01930

Hosna, A., Merry, E., Gyalmo, J., Alom, Z., Aung, Z., & Azim, M. A. (2022). Transfer learning: A friendly introduction. Journal of Big Data, 9(1), 102. https://doi.org/10.1186/s40537-022-00652-w

Kortli, Y., Jridi, M., Al Falou, A., & Atri, M. (2020). Face Recognition Systems: A Survey. Sensors, 20(2), 342. https://doi.org/10.3390/s20020342

Krichen, M. (2023). Convolutional Neural Networks: A Survey. Computers, 12(8), 151. https://doi.org/10.3390/computers12080151
Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. In Advances in neural information processing systems (Vol. 25).
Li, L., Mu, X., Li, S., & Peng, H. (2020). A Review of Face Recognition Technology. IEEE Access, 8, 139110-139120. https://doi.org/10.1109/ACCESS.2020.3011028
Nguyen-Meidine, L. T., Granger, E., Kiran, M., & Blais-Morin, L.-A. (2017, November). A comparison of CNN-based face and head detectors for real-time video surveillance applications. In 2017 Seventh International Conference on Image Processing Theory, Tools and Applications (IPTA) (pp. 1–7). IEEE. https://doi.org/10.1109/IPTA.2017.8310113
Orozco, C. I., Xamena, E., Buemi, M. E., & Berlles, J. J. (2020). Human Action Recognition in Videos Using a Robust CNN–LSTM Approach. Ciencia y Tecnología, (20), 23–36. https://dialnet.unirioja.es/servlet/articulo?codigo=7763841
Parkhi, O. M., Vedaldi, A., & Zisserman, A. (2015). Deep Face Recognition. En Proceedings of the British Machine Vision Conference (BMVC 2015) (pp. 41.1-41.12). https://doi.org/10.5244/C.29.41
Paulchamy, B., Yahya, A., Chinnasamy, N., & Kasilingam, K. (2025). Facial Expression Recognition Through Transfer Learning: Integration of VGG16, ResNet, and AlexNet with a Multiclass Classifier. Acadlore Transactions on AI and Machine Learning, 4(1), 25–39. https://doi.org/10.56578/ataiml040103
Qawaqneh, Z., Mallouh, A. A., & Barkana, B. D. (2017). Deep Convolutional Neural Network for Age Estimation based on VGG-Face Model (No. arXiv:1709.01664). arXiv. https://doi.org/10.48550/arXiv.1709.01664
Sanabria Moyano, J. E., Roa Avella, M. del P., Lee Pérez, O. I., Sanabria Moyano, J. E., Roa Avella, M. del P., & Lee Pérez, O. I. (2022). Tecnología de reconocimiento facial y sus riesgos en los derechos humanos. Revista Criminalidad, 64(3), 61-78. https://doi.org/10.47741/17943108.366
Sanchez, W., & Sigua, E. (s. f.). Reconocimiento facial en video usando Deep learning.
Sarabu, A., & Santra, A. K. (2021). Human Action Recognition in Videos using Convolution Long Short-Term Memory Network with Spatio-Temporal Networks. Emerging Science Journal, 5(1), 25–33. https://doi.org/10.28991/esj-2021-01254
Southwest Minzu University, Key Laboratory of Electronic and Information Engineering, State Ethnic Affairs Commission, Chengdu, 610041, China, Ku, H., Dong, W., & Southwest Minzu University, Key Laboratory of Electronic and Information Engineering, State Ethnic Affairs Commission, Chengdu, 610041, China. (2020). Face Recognition Based on MTCNN and Convolutional Neural Network. Frontiers in Signal Processing, 4(1). https://doi.org/10.22606/fsp.2020.41006
Vimal, C., & Shirivastava, N. (2022). Face and Face-mask Detection System using VGG-16 Architecture based on Convolutional Neural Network. International Journal of Computer Applications, 183(50), 16-21. https://doi.org/10.5120/ijca2022921700
Wang, K. (2024). An Exploration of Face Recognition Methods Based on LBP Algorithm and PCA Analysis. Proceedings of the 2024 AETR Conference on Artificial Intelligence & Emerging Trends in Research and Methodologies. https://madison-proceedings.com/index.php/aetr/article/view/2043
Xie, Y., Wang, H., & Guo, S. (2020). Research on MTCNN face recognition system in low computing power scenarios. Journal of Internet Technology, 21(5), 1463–1475. https://jit.ndhu.edu.tw/article/view/2380/2396
Zhang, K., Zhang, Z., Li, Z., & Qiao, Y. (2016). Joint face detection and alignment using multitask cascaded convolutional networks. IEEE Signal Processing Letters, 23(10), 1499–1503. https://doi.org/10.1109/LSP.2016.2603342
Zhong, Y., Deng, W., Hu, J., Zhao, D., Li, X., & Wen, D. (2021). SFace: Sigmoid-constrained hypersphere loss for robust face recognition. IEEE Transactions on Image Processing, 30, 2587–2598. https://doi.org/10.1109/TIP.2020.3048632