UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS
DEPARTAMENTO DE INGENIERÍA ELÉCTRICA

DISEÑO Y BÚSQUEDA DE ARQUITECTURAS DE REDES NEURONALES CONVOLUCIONALES MEDIANTE NEUROEVOLUCIÓN PARA RECONOCIMIENTO FACIAL CON GRAN VARIACIÓN DE POSE

TESIS PARA OPTAR AL GRADO DE DOCTOR EN INGENIERÍA ELÉCTRICA

JUAN PABLO PÉREZ CABALLERO

PROFESOR GUÍA:
CLAUDIO PÉREZ FLORES

MIEMBROS DE LA COMISIÓN:
PABLO ESTÉVEZ VALENCIA
DOMINGO MERY QUIROZ
GONZALO RUZ HEREDIA

Este trabajo ha sido parcialmente financiado a través del proyecto FONDECYT 1231675 de ANID, además del financiamiento Basal de ANID, AMTC, Proyecto AFB220002, e IMPACT FB210024, y por el Departamento de Ingeniería Eléctrica de la Universidad de Chile.

SANTIAGO DE CHILE
2023

RESUMEN DE LA TESIS PARA OPTAR
AL GRADO DE DOCTOR EN INGENIERÍA
ELÉCTRICA
POR: JUAN PABLO PÉREZ CABALLERO
FECHA: AGOSTO, 2023
PROF. GUÍA: CLAUDIO PÉREZ FLORES

DISEÑO Y BÚSQUEDA DE ARQUITECTURAS DE REDES NEURONALES CONVOLUCIONALES MEDIANTE NEUROEVOLUCIÓN PARA RECONOCIMIENTO FACIAL CON GRAN VARIACIÓN DE POSE

El reconocimiento facial es una técnica biométrica ampliamente utilizada para la autenticación de identidad en varios dominios. A pesar del notable progreso en el campo del reconocimiento de rostros en la actualidad, existen desafíos que se deben abordar, especialmente relacionados al reconocimiento de rostros con gran variación de pose. Siguiendo la idea de las distintas áreas del cerebro especializadas en poses del rostro encontradas en los cerebros humanos y de macacos, en esta tesis se diseñan un conjunto de redes neuronales convolucionales (CNNs) para representar esas regiones, donde cada una responde a un rango particular de orientación de pose. Para esto, se utilizan métodos basados en neuroevolución, empleando algoritmos genéticos (AGs), definiendo la estructura de tres CNNs. Cada CNN evoluciona a través de un AG para un rango particular de orientación de pose, correspondiente a rotaciones pequeñas, medianas y grandes del rostro. Las mejores CNNs obtenidas de cada AG fueron entrenadas con los conjuntos de datos VGGFace2 y MS1M. El rendimiento del método propuesto se evaluó en conjuntos de datos que contienen un número significativo de rostros con gran variación de pose, alcanzando porcentajes de desempeño mayores en comparación al estado del arte: 95,91 %, 95,73 %, 94,60 % y 99,18 % en VGGFace2 (prueba), VGGFace2_FP, CPFLW y CFP_FP, respectivamente.

Agradecimientos

Este trabajo ha sido parcialmente financiado a través del proyecto FONDECYT 1231675 de ANID, además del financiamiento Basal de ANID, AMTC, Proyecto AFB220002, e IMPACT FB210024, y por el Departamento de Ingeniería Eléctrica de la Universidad de Chile.

Tabla de Contenido

Índice de Tablas

Índice de Ilustraciones

Capítulo 1

Introducción

1.1. Motivación

1.1.1. Reconocimiento Facial

El Reconocimiento Facial (en inglés *Face Recognition*, FR) ha sido una técnica biométrica ampliamente utilizada para la autenticación de identidad en varios dominios [6, 7, 8]. Los recientes avances en *Deep Learning* y el desarrollo de nuevos modelos han producido mejoras significativas en el desempeño de los sistemas de FR [6, 9, 10]. A pesar del notable progreso en el campo del FR, todavía existen desafíos que deben ser abordados. En los últimos años, varios estudios han destacado los desafíos restantes en FR, que incluyen el reconocimiento con grandes variaciones de pose en aplicaciones de vigilancia [11, 12]; la recuperación incorrecta causada por imágenes de fondo no etiquetadas [13]; la pérdida de información discriminativa en imágenes sintetizadas por ciertas funciones de pérdida [14] ; el rendimiento degradado con rostros desalineados [15] y las dificultades en el entrenamiento de redes neuronales convolucionales (CNNs) con conjuntos de datos pequeños [16]. Estos desafíos requieren una investigación adicional para superar las limitaciones de los sistemas FR actuales. En particular, investigaciones recientes han mostrado que el FR con grandes variaciones de pose para aplicaciones de vigilancia ha producido pobre desempeño en conjuntos de datos en los que los rostros frontales no son predominantes [6, 17]. Los métodos del estado del arte (SOTA) han logrado una exactitud de casi el 99.9 % en conjuntos de datos como LFW [13] [18], y Megaface [19, 20], que consisten principalmente en rostros frontales o pocas variaciones de pose. Sin embargo, los conjuntos de datos con rostros que incluyen variaciones en edades [21], poses [22, 23], iluminación [24], sensores o estilos, aún no han alcanzado altos niveles de porcentaje de acierto [6]. Estos desafíos destacan la necesidad de una investigación adicional

para mejorar el rendimiento de los métodos FR en escenarios donde los rostros no son predominantemente frontales. Además, el aumento del uso de drones de vigilancia y monitoreo ha presentado nuevos desafíos en el FR y la clasificación de acciones en condiciones de ángulos variables, efectos de movimiento, cambios extremos de pose y variaciones de iluminación [25, 26, 27]. Los autores en [28] demostraron que varios métodos producen una reducción en porcentaje de acierto de más del 10 % al cambiar la verificación de coincidencia de rostros frontal-frontal a frontal-perfil. El reconocimiento facial cruzado de poses sigue siendo una tarea extremadamente desafiante [29, 30, 31]

1.1.2. Antecedentes biológicos de FR con variaciones de pose

En complemento a lo anterior, estudios en neuroimagen y electrofisiología han explorado las regiones cerebrales involucradas en el procesamiento de la información sobre la orientación del rostro y el cuerpo en humanos y macacos [1]. Los estudios de neuroimagen en humanos han revelado diferentes patrones de respuesta a diferentes orientaciones del rostro en varias regiones, como el área occipital sensible al rostro (OFA), el área fusiforme sensible al rostro (FFA) y el surco temporal superior posterior (pSTS), así como el complejo lateral occipital sensible a objetos (LOC) y la corteza visual temprana [32]. De manera similar, los estudios en macacos han demostrado que las neuronas en las zonas del rostro lateral media y media fundus sensible a la cara, y el surco temporal superior anterior (aSTS), responden a orientaciones específicas [33].

En los cerebros de primates, típicamente hay regiones en la corriente visual ventral que responden selectivamente a los rostros. En macacos, estas áreas se encuentran en partes similares de la corteza inferotemporal en diferentes individuos, aunque la correspondencia con características anatómicas específicas no se había reportado previamente. En [1] se concluye que es posible determinar la existencia de áreas de la corteza lateral temporal en macacos que son selectivamente responsivas a rostros. Los arreglos de electrodos implantados verificaron que estos estímulos contienen neuronas selectivas para rostros. En la Figura 1.1 se puede apreciar de manera gráfica los resultados de estos experimentos.

Estos hallazgos proporcionan información sobre los mecanismos neurales subyacentes al procesamiento de la orientación facial y del cuerpo en humanos y monos, mostrando que varias regiones del cerebro están especializadas en responder a diferentes orientaciones del rostro [1, 32, 33]. Los resultados descritos de los mecanismos neurales en humanos y monos macacos utilizados para procesar la orientación facial [1, 32, 33], proporcionan una posible línea de investigación para abordar la necesidad de modelos mejorados en el área de FR con grandes variaciones de pose. Nuestro trabajo sigue esta línea de investigación. Se pueden

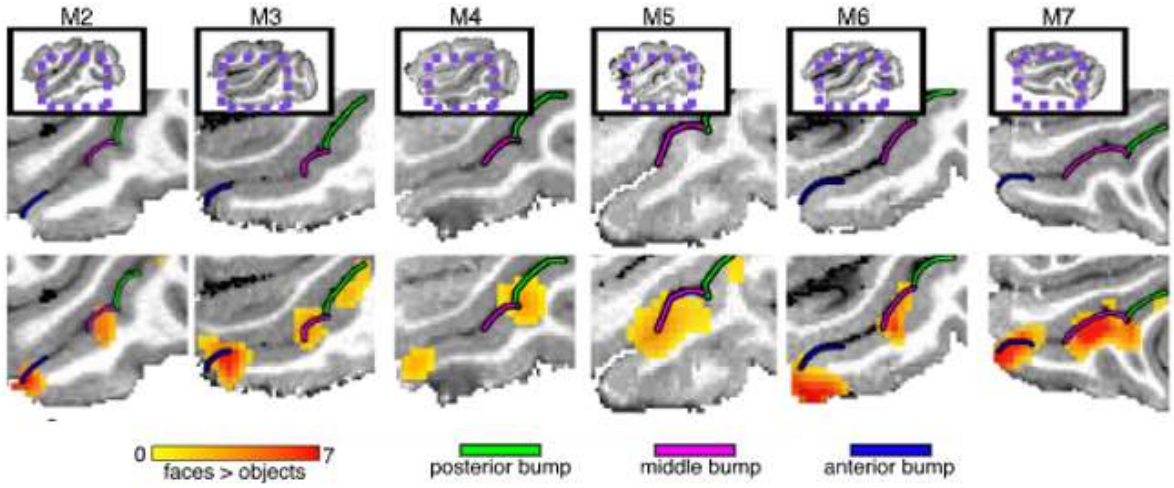


Figura 1.1: La localización anatómica de la selectividad de rostros en el STS (surco temporal superior) del hemisferio derecho fue identificada en la parte inferior del surco temporal superior en 6 monos criados con experiencia normal de rostros, mostrando actividad preferencial para rostros en comparación con objetos. Extraída de [1].

desarrollar nuevas arquitecturas de modelos de redes neuronales especializadas en un rango de poses del rostro específico siguiendo la idea de tener varias zonas del rostro en el cerebro que respondan a diferentes orientaciones. Una posible solución que proponemos en este trabajo para FR con grandes cambios de pose es proporcionar varias redes neuronales convolucionales, como regiones del rostro, donde cada una está ajustada a un rango de orientación específico, siendo esta una idea original.

En el presente estudio, proponemos un nuevo enfoque para mejorar la exactitud del reconocimiento facial, incluyendo aquellos con grandes variaciones de pose.

1.2. Hipótesis

- Es posible modificar la arquitectura de una CNN, a través de un algoritmo genético, para tener un reconocimiento de rostros mejorado en un rango específico de pose del rostro. El desempeño de la CNN estructurada para un rango de ángulos específico debe ser superior a las redes sin estructurar que fueron entrenadas para todos los ángulos de rostros.

- Es posible obtener resultados que superen el estado del arte en reconocimiento de rostros con grandes cambios de pose, simulando lo que se ha observado en el cerebro de humanos y macacos con las regiones faciales incluyendo áreas especializadas en rangos

de pose, mediante la combinación de varias CNNs, en que cada una está sintonizada a un rango de pose específico.

- Las arquitecturas de CNNs obtenidas para cada rango de pose deberían presentar diferencias en términos de los parámetros que definen cada arquitectura y la capacidad de reconocimiento facial para los distintos rangos de pose.

1.3. Objetivo General

Desarrollar un método de reconocimiento facial que permita identificar rostros humanos con variaciones de pose significativas utilizando un modelo original basado en varias redes neuronales convolucionales estructuradas, utilizando algoritmos genéticos, para tener buen desempeño en distintos rangos de pose del rostro.

1.4. Objetivos Específicos

- Desarrollar un método para modificar la arquitectura de una CNN, a través de un algoritmo genético, para que obtenga buenos resultados de reconocimiento facial en un rango específico de rotación del rostro: pequeñas rotaciones ($\pm 10°$), rotaciones medianas ($\pm[10°, 40°]$) y grandes rotaciones ($\pm[40°, 90°]$). Se utilizarán algoritmos genéticos de dos niveles (2LGA).

- Desarrollar un método de reconocimiento facial que integre varias CNNs evolucionadas para tener una respuesta mejorada a rangos de poses faciales específicas, utilizando un detector de pose de rostros.

- Medir desempeño de este nuevo método de reconocimiento facial, usando bases de datos internacionales especialmente diseñadas para medir la tarea de reconocimiento con variaciones de pose, comparando los resultados obtenidos con el estado del arte.

- Identificar y visualizar las diferencias en las arquitecturas de las CNN evolucionadas, en términos de los parámetros clave que definen cada arquitectura y su rendimiento en el reconocimiento facial en distintos rangos de pose.

1.5. Principales Contribuciones

La primera contribución de nuestro nuevo enfoque es crear varias CNNs siguiendo la idea de implementar diversas regiones faciales para el reconocimiento de rostros, en las cuales cada una está especializada y se aplica en un rango de orientación facial particular. La segunda contribución es emplear neuroevolución, utilizando algoritmos genéticos (GAs), para crear y evolucionar diferentes CNNs, en las que cada una está especializada en un rango de pose facial particular. Los GAs descubrirán arquitecturas casi óptimas que pueden mejorar la exactitud de FR con grandes variaciones de pose de manera eficiente. Para lograr esto, se realizarán varios experimentos, utilizando conjuntos de datos estándar, que permitirán comparar nuestros resultados con los del estado del arte. Se propone desarrollar CNNs para tres rangos de pose facial diferentes: pequeñas rotaciones, de -10° a 10°; rotaciones medianas, de -10° a -40° y de 10° a 40°; y grandes rotaciones, de -40° a -90° y de 40° a 90°, utilizando tres modelos distintos obtenidos con algoritmos genéticos entrenados con las distintas poses faciales para cada rango de pose particular.

Este investigación culminó en la publicación de un artículo en IEEE Access, revista indexada en Web of Science (WoS) [34]. Esta publicación destaca la validez y relevancia científica de nuestra investigación, contribuyendo significativamente al conocimiento en el campo del reconocimiento facial y reflejando el estándar académico del trabajo.

1.6. Estructura de la Tesis

El documento está organizado de la siguiente manera: El capítulo 2 muestra el estado del arte y el trabajo relacionado, el capítulo 3 describe el método desarrollado, los experimentos que fueron diseñados para demostrar el desempeño del método y las bases de datos utilizadas en el estudio. El capítulo 4, correspondiente a Resultados, análisis y discusión presenta los resultados obtenidos y, finalmente, el 5 presenta las conclusiones derivadas del trabajo realizado, resumiendo los hallazgos y resultados obtenidos. Además, se mencionan las posibles líneas de investigación futura que pueden derivarse de esta investigación.

Capítulo 2

Estado del Arte

2.1. Reconocimiento de rostros con variación de pose

Varios estudios de estado del arte [6, 7, 35] resaltan los desafíos en curso en el reconocimiento facial, tal como se evidencia en conjuntos de datos no saturados, es decir, que en esos conjuntos aún se tiene margen de mejora en términos de tasa de reconocimiento del modelo. Por ejemplo, tres conjuntos de datos importantes, MegaFace, MS-Celeb-1M y IJB-A/B/C, plantean desafíos para el reconocimiento facial a gran escala con un gran número de candidatos, el reconocimiento facial de baja/una sola muestra y el reconocimiento facial con una gran variabilidad de poses. Si bien los métodos de estado del arte logran una gran exactitud en ciertos conjuntos de datos, como LFW y Megaface, no funcionan tan bien en otros conjuntos de datos, como CFP_FP y CPLFW. Aún existen desafíos fundamentales, como el balance de rostros entre diferentes edades, poses, sensores o estilos [6]. Para mejorar el FR, es crucial abordar problemas como los sesgos de raza, género, pose y edad que están presentes en el FR profundo [6, 35]. Aunque el FR profundo ha tenido éxito en aplicaciones que requieren de cooperación de los usuarios, lograr un reconocimiento universal en todas las configuraciones sigue siendo un objetivo ambicioso [6, 35]. En la práctica, recopilar y etiquetar muestras suficientes para numerosas escenas del mundo real es un desafío. Una posible solución es desarrollar un modelo general que se pueda transferir a escenas específicas de aplicaciones [35]. Aunque la adaptación profunda de dominios se ha aplicado recientemente para reducir el sesgo del algoritmo en escenas y personas de diferentes razas, encontrar una solución general para transferir el FR sigue siendo en gran medida un tema abierto [6, 35]. Se requiere investigación adicional para abordar estos desafíos y lograr el FR en diversas configuraciones [6, 7, 35]. A pesar de los avances en el FR, el rendimiento de los modelos de FR a menudo no cumple con los requisitos de las aplicaciones del mundo real. Para abor-

dar este problema, se han dedicado esfuerzos significativos para diseñar nuevos métodos que puedan abordar problemas específicos relacionados con datos limitados en escenas realistas [6, 35]. Estos problemas incluyen reconocimiento facial con variación de pose, edades, baja resolución de las imágenes o poca cantidad de muestras [6, 7, 36, 35]. Mejorar la exactitud y eficacia de los modelos de FR en estos escenarios es fundamental para lograr un FR universal en diversas aplicaciones y escenas [6, 7, 36, 35]. Aunque ha habido algunos estudios previos sobre FR utilizando Búsqueda de Arquitecturas Neuronales (en inglés *Neural Architecture Search*, NAS), no se centran explícitamente en resolver el problema de la variación de poses. Además, estos estudios han realizado pruebas principalmente en conjuntos de datos con una variación de poses limitada, como LFW [37, 38].

A lo largo de los años, se han desarrollado numerosos algoritmos diseñados para la tarea de reconocimiento de rostros con variaciones de pose. En [39], se presenta un modelo probabilístico elástico de partes (PEP). El método captura la distribución de la apariencia espacial de las características del rostro de todas las imágenes faciales, y posteriormente construye una representación concatenando diferentes descriptores en secuencia. En [40], se intenta realizar la frontalización en el espacio de características, en lugar de aplicarlo en el espacio de la imagen. Un bloque de mapeo residual equivariante profundo (DREAM) añadió residuos dinámicamente a una representación de entrada para transformar una cara de perfil a una imagen frontal. En [41], se propuso combinar la extracción de características con el aprendizaje de subespacios de múltiples vistas para hacer que las características sean simultáneamente más resistentes y discriminativas a la pose.

En [8], un método llamado Modelo Invariante a la Pose realiza la frontalización del rostro y aprende representaciones invariantes a la pose, tanto en conjunto como de extremo a extremo, y además introdujo el entrenamiento adversario no supervisado entre dominios y una estrategia de aprendizaje para proporcionar imágenes de referencia frontales de alta fidelidad. En [42], se presentó un enfoque novedoso para la frontalización de rostros utilizando una red generativa de adversarios de identificación de múltiples dominios y fusión de características pre-entrenadas (PM-GAN). En este método, las características pre-entrenadas obtenidas de conjuntos de datos a gran escala se fusionan con las características originales del codificador para mejorar la diversidad y la robustez de dichas características. En [43] se presenta un enfoque novedoso llamado Redes de Atención Eficientes y Ligeras (ELANet) que aborda el problema de la pose y la edad en la precisión de los sistemas FR. El enfoque propuesto aprovecha la importancia de regiones locales similares en entornos donde la apariencia y la geometría del rostro sufren cambios significativos.

Relacionado a las CNNs livianas, los autores de [70] introducen una nueva familia de modelos ligeros de reconocimiento facial conocidos como ConvFaceNeXt, motivados por Conv-

NeXt y MobileFaceNet, que buscan mejorar la eficiencia en términos de operaciones de punto flotante (FLOPs), parámetros y tamaño del modelo. Según los resultados experimentales, el modelo ConvFaceNeXt propuesto logra resultados competitivos o incluso mejores en comparación con modelos previos de reconocimiento facial ligeros, con un número significativamente menor de FLOPs, parámetros y tamaño del modelo.

Los autores de [44], sostienen que los métodos convencionales tienen dificultades para reflejar la distribución completa de los conjuntos de datos, porque un minibatch de pequeño tamaño contiene solo una pequeña porción de todas las identidades. Para superar esta dificultad, proponen un nuevo método llamado BroadFace, que es un proceso de aprendizaje para considerar un conjunto masivo de identidades de manera exhaustiva. BroadFace logra resultados de vanguardia con mejoras significativas en nueve conjuntos de datos en tareas de verificación facial 1:1 e identificación facial 1:N, y también es efectivo en la tarea de recuperación de imágenes.

En [45] se propone un método llamado MagFace para aprender características unificadas para el reconocimiento facial. Al alejar las muestras ambiguas de los centros de clase, MagFace mejora la distribución de características dentro de la clase de los trabajos previos basados en margen para el reconocimiento facial. Los resultados teóricos y experimentales adecuados demuestran que MagFace puede trabajar con la calidad de la imagen facial de entrada. Como marco general, MagFace puede extenderse potencialmente para beneficiar otras tareas de clasificación, como el reconocimiento de objetos detallados o la reidentificación de personas.

En [46] se propone el Aprendizaje de Prototipos Variacionales (VPL), que representa cada clase como una distribución, en lugar de un punto en el espacio latente. VPL se puede integrar directamente en métodos de softmax mejorando el rendimiento del reconocimiento facial profundo.

En [47], los autores, inspirados en una operación de convolución deformable, desarrollaron una red de convolución llamada Deformable Face Net (DFN), diseñada para aprender la alineación a nivel de características y la extracción de características simultáneamente para FR. En [48] se propuso un marco novedoso no supervisado, llamado rotar y renderizar, que puede sintetizar caras rotadas fotorrealistas utilizando solo colecciones de imágenes de una sola vista del rostro. Los autores sostienen que su marco puede actuar naturalmente como un motor de aumento de datos efectivo para impulsar los sistemas FR.

En [6] se hace un resumen del estado del arte en reconocimiento de rostros utilizando aprendizaje profundo, y se presentan los mejores resultados de reconocimiento para diversos métodos de estado del arte (SOTA, por sus siglas en inglés), utilizando los conjuntos de datos de prueba ampliamente utilizados para FR con variaciones de pose. En [30], los autores intro-

dujeron la pérdida ElasticFace, que mejoró los resultados de ArcFace y CosFace al mejorar la estimación de error entre características en los entrenamientos. En [49], se propuso un nuevo enfoque para adaptar las funciones de pérdida al introducir la calidad de la imagen como factor. Los autores argumentaron que la ponderación de las muestras clasificadas incorrectamente debería considerar la calidad de la imagen, teniendo en cuenta la complejidad relativa de las muestras fáciles y difíciles. Se propuso una nueva función de pérdida, AdaFace, que utiliza una función de margen adaptativo para aproximar la calidad de la imagen mediante normas de características. Los resultados utilizando AdaFace superaron a muchos trabajos del SOTA.

En [50] se propone un nuevo método de alineación facial para la tarea de FR invariante a la pose, llamado alineación adaptativa de pose (APA), que aprende plantillas de alineación según las poses faciales para reducir las diferencias intraclase y corregir el ruido. También se introduce un método de normalización de características para generar una representación de características más discriminativas. El método propuesto muestra un buen rendimiento en conjuntos de datos de reconocimiento facial de última generación, incluidos los conjuntos de datos IJB-A, IJB-C y CPLFW.

En [51], se propone un modelo de Redes Generativas Adversarias de Pose Cruzada (CP-GAN) para el FR invariante a la pose, aprendiendo a mapear caras de perfil a caras frontales con la misma identidad. CP-GAN utiliza un generador de red U-net codificador-decodificador y una red discriminativa Siamesa para la extracción profunda de características. Se combina *GAN loss*, con funciones de pérdidas de simetría, de regiones y de identidad para lograr la frontalización preservando la identidad.

En [52], se propone un método basado en Redes Generativas Adversarias (GAN) que preserva la estructura geométrica (GSP-GAN) para la frontalización y reconocimiento de rostros con múltiples poses. El modelo utiliza un generador autoencoder con pérdida de percepción y pérdida de la norma L1 para sintetizar una imagen de cara frontal con la misma identidad que la imagen de entrada. Se utiliza un bloque de autoatención para preservar la estructura geométrica del rostro en el discriminador. GSP-GAN supera a los modelos de última generación en los conjuntos de datos Multi-PIE, LFW y CFP.

En [53], se propone un marco de entrenamiento alternado conjunto (JAT) para el FR de cola larga, que utiliza tanto datos de cola larga como datos de cola con muestreo equilibrado de instancias y de clases. Se aplica aumentación de datos (en inglés *Data Augmentation*) y mezcla basada en el margen para compensar la falta de muestras y la falta de variaciones intraclase. La combinación propuesta de estrategias logra un rendimiento de última generación en 8 conjuntos de datos de rostros.

En [37] se propone una búsqueda de arquitectura de redes neuronales profundas para el FR, combinando NAS y el aprendizaje por refuerzo (en inglés o *Reinforcement Learning*). El método propuesto optimiza NAS incorporando la latencia de evaluación en las recompensas y utiliza el algoritmo basado en gradientes (en inglés, *Policy Gradient Methods*) para buscar arquitecturas automáticamente. Las arquitecturas de red resultantes logran una exactitud de última generación en los conjuntos de datos MS1M y LFW, que tienen un tamaño de red relativamente pequeño.

En [54], los autores presentan PocketNet, una solución ligera de FR que utiliza NAS. Proponen un enfoque de entrenamiento novedoso llamado destilación de conocimiento en múltiples etapas (KD), donde el conocimiento del modelo maestro se destila al modelo alumno en diferentes etapas de entrenamiento. Su red más pequeña, PocketNetS-128, de solo 0.92 millones de parámetros, logra resultados altamente competitivos.

En [55], se indica que el FR de última generación logra una alta exactitud al depender de imágenes web recopiladas. Para disminuir los sesgos de la base de datos, introducen un conjunto de datos sintético a gran escala. El aumento de datos reduce la brecha entre lo sintético y lo real, lo que resulta en una reducción del 52.5 % en la tasa de error en LFW en comparación con SynFace [56].

En [57], se propone el Generador de Caras con Condiciones Duales (DCFace), utilizando un modelo de difusión para controlar la apariencia del sujeto y los factores externos. Los modelos de reconocimiento facial entrenados en imágenes sintéticas de DCFace superan los resultados de última generación en términos de porcentaje de aciertos en tareas de verificación.

En [58], se presenta un enfoque novedoso para el problema de la distribución de dominios de cola larga en el reconocimiento facial, en respuesta al hecho de que solo un pequeño número de dominios aparecen con frecuencia, mientras que otros dominios existen en menor medida. El estudio propone un método llamado mecanismo de equilibrio de dominios (DB), que incluye un indicador de frecuencia de dominios (DFI) para identificar dominios principales y de cola; un bloque de mapeo de equilibrio residual (RBM) para equilibrar la distribución de dominios ajustando la red; y un margen de equilibrio de dominio (DBM) para optimizar el espacio de características de los dominios de cola. Los experimentos en varios conjuntos de datos de reconocimiento facial muestran que el método propuesto mejora significativamente la generalización y logra un rendimiento mejorado.

En 2021, se publicó un nuevo conjunto de datos de referencia para el reconocimiento facial a gran escala [59]. El conjunto de datos WebFace260M consta de 4 millones de identidades y 260 millones de rostros, proporcionando además una herramienta excelente para la limpieza

y el reconocimiento facial profundo de un millón de clases [59]. Los autores implementaron el pipeline CAST (limpieza automática mediante autoentrenamiento) para limpiar automáticamente el ruidoso WebFace260M y obtuvieron un conjunto de entrenamiento limpio llamado WebFace42M. Se entrenaron modelos ArcFace [60] utilizando los subconjuntos WebFace42M, WebFace12M y WebFace4M, que superaron todos los FR en el desafiante conjunto de datos IJB-C [61]. ArcFace también se comparó con los conjuntos de datos MS1M [62], IMDB_Face [63] y MegaFace2 [64], donde se obtuvo un desempeño menor en comparación a WebFace42M.

Varios autores han reconocido que el diseño manual de arquitecturas CNN por expertos humanos es excesivamente lento [65]. La estructura de una CNN puede ser muy diferente dependiendo de la tarea para la cual esté diseñada. Por ejemplo, la arquitectura de una CNN para reconocimiento de iris [66] es diferente a la utilizada para detección de objetos o segmentación semántica [67]. Existe un gran interés en automatizar esta tarea, es decir, encontrar el tipo de operación implementada en cada capa (convolución, operaciones de agrupación, capas densas, etc.) y los hiperparámetros asociados a la operación (número de filtros, tamaño del kernel y pasos para una capa convolucional, etc.), utilizando los conjuntos de datos de entrenamiento y prueba disponibles [68, 69]. Uno de los métodos con mayor potencial en el diseño de CNN, NAS [68, 70], ha superado a las arquitecturas diseñadas manualmente y comúnmente utilizadas en tareas como clasificación de imágenes [71, 72, 73] y segmentación semántica [74, 75]. Como se plantea en la hipótesis, entonces la neuroevolución puede servir para diseñar automáticamente CNNs que tengan un rendimiento mejorado en rostros dentro de un rango de orientación de pose específico.

2.2. Búsqueda de Arquitectura Neural

La Neuroevolución es un subtema de NAS, que consiste en buscar arquitecturas de CNNs y NNs automáticamente, utilizando varios métodos como el Aprendizaje por Reforzamiento (RL), la Optimización basada en Gradientes (GD), la Optimización Bayesiana (BO) y los métodos evolutivos. La Neuroevolución podría impactar fuertemente en el desarrollo de tecnologías futuras basadas en *Deep Learning* de dos maneras. La primera es permitir a los investigadores encontrar mejores arquitecturas de CNN para tareas específicas, reduciendo así el tiempo y el esfuerzo requeridos para encontrarlas. La segunda es que la Neuroevolución permite a los no expertos encontrar modelos de CNN cercanos al óptimo. La Neuroevolución podría ser utilizada por investigadores de otros campos sin requerir una experiencia avanzada en DL. Por ejemplo, la Neuroevolución se ha aplicado con éxito a imágenes médicas [76, 77, 78], reconocimiento de voz [79, 80], reconocimiento de emociones,[81], clasificación de escenas [82], entre otros [83, 84, 85].

El espacio de búsqueda de NAS determina la estructura de la red durante el proceso de búsqueda, con el objetivo de asegurar un modelo de muestreo racional y de alta calidad. Aunque introducir conocimiento previo para tareas específicas puede reducir el espacio de búsqueda, también limita la generación de redes más allá de la experiencia previa [86]. La estructura base de la red NAS comprende tres tipos de espacio de arquitectura: espacio de búsqueda basado en cadenas (*Chain-Based Search Space* en inglés), espacios de búsqueda de múltiples ramas y basados en bloques. Su espacio de operadores consiste en convoluciones, pooling, conexiones residuales y otras estructuras topológicas [86].

La estrategia de búsqueda de NAS determina qué algoritmo es capaz de identificar las configuraciones óptimas de la arquitectura de la red de manera rápida y precisa [86]. Dependiendo de la estrategia de búsqueda específica utilizada, NAS se puede clasificar en cuatro categorías diferentes [37]: Métodos basados en aprendizaje por refuerzo, como NASNet, BlockQNN, ENAS y MnasNet; Métodos basados en aprendizaje evolutivo, como Hierarchical y AmoebaNet; Métodos continuos diferenciables, como DARTS, NAO y DSO; Otros métodos de búsqueda que incluyen SMASH, PNAS, Auto-Keras, Graph hypernet, Proxyless NAS y Efficient Multi-Scale Architectures [87, 88, 71, 89, 90, 91, 92].

Dado el significativo tiempo y recursos computacionales requeridos por NAS, las estrategias de aceleración comúnmente utilizadas incluyen el uso de compartición de parámetros, morfismo de red y poda de red [37, 86]. Sin embargo, la elección de los indicadores de evaluación del rendimiento para los algoritmos de NAS varía según la tarea y el escenario en cuestión [86]. Los indicadores típicos incluyen precisión de prueba, número de parámetros, latencia de inferencia y utilización de memoria.

Los algoritmos genéticos se inspiran en los procesos de evolución y selección natural y aprovechan operadores genéticos para obtener soluciones de alta calidad. Los elementos principales de los procedimientos de AG incluyen la inicialización, evaluación de *fitness*, operadores de población (mutación y cruce) y selección. Se utiliza el concepto de reproducción para englobar los procesos de selección, mutación y cruce. Los algoritmos genéticos son apropiados para buscar problemas de arquitecturas de CNN ya que son libres de gradientes y no son sensibles a los mínimos locales. Además, los AG tienen una buena capacidad de búsqueda global en comparación con los métodos de optimización de descenso de gradiente, que pueden converger a un mínimo local en casos de superficies no convexas. Además, pueden manejar problemas incluso cuando no se dispone de una función objetivo explícita o exacta, lo cual es el caso del problema de búsqueda de arquitectura de CNN. Por lo tanto, los AG son una herramienta útil para abordar problemas de búsqueda y optimización en general, y son especialmente adecuados para buscar arquitecturas de CNN debido a sus ventajas en términos de búsqueda global y capacidad de manejar problemas sin una función objetivo explícita.

Recientemente, se desarrolló un Algoritmo Genético (GA) para NAS y se aplicó a varios problemas de clasificación de imágenes [65]. Este estudio utilizó conjuntos de datos como MNIST Variants (5 conjuntos) [55][93], Fashion MNIST [94] y CIFAR-10 [95]. El GA propuesto en [65] utilizó un enfoque de dos niveles, que incorpora una evaluación rápida de la población mediante períodos de entrenamiento más cortos, seguida de un entrenamiento más prolongado de los mejores individuos en generaciones seleccionadas.

Capítulo 3

Metodología

Se propone un enfoque novedoso para mejorar la exactitud en el reconocimiento de rostros en grandes conjuntos de datos que poseen grandes variaciones de pose. El nuevo enfoque se basa en la idea de crear estructuras especializadas para distintas orientaciones del rostro, como las presentes en los cerebros de humanos y macacos, para mejorar el rendimiento de FR para un rango de orientación de pose facial específico. Para esto, se propone diseñar y optimizar arquitecturas de CNNs cercanas para varios rangos de pose facial utilizando Neuroevolución con GAs para crear estas estructuras o *patches* faciales especializados para ciertos rangos de pose. Para entrenar cada una de las CNNs que conforman la población del algoritmo, se definieron tres categorías de rango de pose en función de las orientaciones del ángulo facial: pequeñas rotaciones, desde -10° hasta 10°; rotaciones medias, desde -10° hasta -40° y desde 10° hasta 40°; y rotaciones grandes, desde -40° hasta -90° y desde 40° hasta 90°. Se utilizaron tres GAs para buscar arquitecturas de CNNs, una para cada una de las tres categorías de rango de pose, con 20 individuos en cada población inicial. [65]. Esta clasificación permite que las CNNs se especialicen en rangos específicos de rotación, lo que probablemente mejora su precisión y eficiencia en la tarea de reconocimiento o clasificación facial. Los AG son técnicas de optimización y búsqueda inspiradas en la evolución natural. Se utilizan para encontrar soluciones óptimas en problemas complejos. En este contexto, se emplean para buscar las mejores arquitecturas de CNN para cada categoría de rango de pose. Los AG son adecuados para este propósito porque pueden explorar eficientemente un amplio espacio de soluciones, el uso de NAS, como se menciona en los antecedentes, es una estrategia efectiva para encontrar arquitecturas de CNN que sean óptimas para tareas específicas, en este caso, la clasificación de imágenes basada en diferentes rangos de pose facial.

La función de *fitness* se definió como el desempeño de reconocimiento de la arquitectura de cada individuo CNN, y la búsqueda se ejecutó durante más de 54 generaciones para cada

uno de los GAs. Se realizaron experimentos en el conjunto de datos VGGFace2, un conjunto de datos estándar para FR, que contiene una gran cantidad de imágenes con diversas poses y condiciones de iluminación. Primero, se dividió el conjunto de datos VGGFace2 en conjuntos de entrenamiento y prueba, tal como se ha utilizado ampliamente en la literatura [4].

Se empleó un GA de dos niveles (2LGA) [65] para la optimización de las arquitecturas de CNNs con orientaciones específicas de rango facial. El primer nivel del GA evalúa muchos individuos rápidamente (entrenamiento corto), mientras que el segundo nivel evalúa solo aquellos con los mejores resultados más finamente (entrenamiento largo) [65]. El espacio de búsqueda tiene pocas restricciones, lo que permite el uso de arquitecturas de CNNs de diferentes tamaños, formas y conexiones de salto entre nodos. El 2LGA se aplicó previamente a problemas de reconocimiento de patrones en varios conjuntos de datos, como cinco variantes del dataset MNIST, Fashion-MNIST y CIFAR-10, logrando resultados significativamente mejores que los del SOTA [65]. Como en [65], se utilizó un espacio de búsqueda más versátil que permite el uso de arquitecturas de diferentes formas y longitudes. Nuestro espacio de búsqueda comprende conexiones de salto entre operaciones, parámetros de operación y hiperparámetros de la red. Nuestra estrategia de codificación, para incorporar los hiperparámetros en el Algoritmo Genético, fue utilizada previamente [65] con buenos resultados en conjuntos de datos como MNIST y CIFAR. En la Tabla 3.1 se muestran los principales parámetros y los valores que estos pueden adoptar para generar un individuo válido. Esto permite al GA buscar soluciones con menos sesgo humano.

Tabla 3.1: Resumen de la información codificada en el espacio de búsqueda propuesto.

Descripción		Dominio de Valores
Parámetros estructurales	tasa de crecimiento	[1.5, 4.5]
	número de celdas	{1, 2, 3, 4}
Hiperparámetros	Warmup	[0, 0.5]
	tasa de aprendizaje	[0.002, 0.125]
Parámetros del nodo	Operación	{Convolución, Identidad, Maxpool}
Parámetros de Convolución	método de unión	{CAT, SUM}
	factor de mapa de características	[0.1, 0.2]
	Activación	{ReLU, ELU, PReLU}
	Dropout	[0, 0.6]
	tamaño de kernel	{1, 3, 5}
Parámetros de Maxpool	tamaño de kernel	{2, 3, 5}

El parámetro de tamaño de *batch*, que determina el número de ejemplos de entrenamiento

en cada paso hacia adelante/atrás, se eligió en función de la memoria de la GPU disponible. Se utilizó un tamaño de *batch* de 64 imágenes en nuestro estudio. Los modelos de CNNs se entrenaron desde cero utilizando un optimizador de descenso de gradiente estocástico. Se empleó un programa de tasas de aprendizaje con un valor inicial de 0.1 y un factor de disminución de 0.1. La función de pérdida utilizada ara todos los modelos fue ArcFace [96].

La función de pérdida ArcFace es una función de pérdida popular utilizada en tareas de reconocimiento facial [96]. Es una variante de la función de pérdida softmax diseñada para aumentar el margen angular entre diferentes clases en el espacio de características. En ArcFace, la salida de la última capa completamente conectada se normaliza primero para tener una longitud unitaria. Luego, se utiliza y aprende una matriz de pesos W para cada clase, de manera que el producto punto entre el vector de características normalizado y la matriz de pesos de la clase correcta se maximice, mientras que se minimiza el producto punto con las matrices de pesos de las otras clases [48]. La función coseno se utiliza como métrica de distancia entre el vector de características normalizado y la matriz de pesos [96].

El margen m y el factor de escala s son hiperparámetros que controlan el margen angular y la escala de los vectores de características de salida, respectivamente [96]. El margen m aumenta la separación entre las diferentes clases en el espacio de características, mientras que el factor de escala s controla el tamaño de los vectores de características [96]. Se expresa matemáticamente como:

$$L = -\frac{1}{N}\sum_{\mathrm{i}=1}^{N}\log\frac{\mathrm{e}^{s(\cos(\theta y_{\mathrm{i}}+m))}}{\mathrm{e}^{s(\cos(\theta y_{\mathrm{i}}+m))}+\sum_{j=1,j\neq y_{\mathrm{i}}}^{n}\mathrm{e}^{s\cos\theta_j}}. \tag{3.1}$$

Donde:

- L es la función de pérdida (loss function).
- N es el número de muestras en un lote.
- s es una escala que se aplica a los logits (es decir, la salida de la última capa de la red neuronal antes de la activación softmax).
- $\theta_{y_{\mathrm{i}}}$ es el ángulo entre la característica extraída y el peso de la clase verdadera (la clase correcta a la que pertenece la imagen).
- m es un margen que se añade al ángulo $\theta_{y_{\mathrm{i}}}$ para incrementar la discriminación entre clases.
- y_{i} es la etiqueta de la clase verdadera para la i-ésima muestra.

- n es el número total de clases.

La idea detrás de ArcFace es que al añadir un margen angular m al ángulo θ_{y_i} entre las características y el peso de la clase verdadera, se fuerza a que las características de las diferentes clases estén más separadas en el espacio angular. Esto mejora la capacidad del modelo para distinguir entre diferentes identidades, incluso cuando las variaciones intraclase (diferencias dentro de la misma clase) son significativas.

ArcFace modifica la función de pérdida de manera que promueve una mayor separación angular entre las diferentes clases en el espacio de las características, lo que lleva a una mejor discriminación en tareas de reconocimiento facial.

La función de pérdida ArcFace [96] se puede optimizar utilizando algoritmos de descenso de gradiente [48]. Se ha demostrado que logra un rendimiento de última generación en varios conjuntos de datos de reconocimiento facial, superando a otras funciones de pérdida populares como softmax y triplet loss [6, 9].

Tanto la función de pérdida como las funciones de activación utilizadas en este estudio están definidas y demostradas en las referencias proporcionadas.

Las funciones de activación consideradas fueron las siguientes: ReLU (Unidad Lineal Rectificada): ReLU es una función de activación comúnmente utilizada en el aprendizaje profundo [97, 98]. Toma una entrada x y devuelve el máximo entre 0 y x, como se muestra a continuación,

$$\text{ReLU}(x) = \text{máx}(0, x). \tag{3.2}$$

Es una función lineal por partes que ofrece un buen rendimiento en muchos modelos de aprendizaje profundo. Sin embargo, ReLU puede sufrir del problema de “ReLU muerta", en el cual algunas neuronas se vuelven inactivas y producen una salida de cero para todas las entradas [97, 98].

ELU (Unidad Lineal Exponencial): ELU es una función de activación diseñada para superar el problema de “ReLU muerta"[97, 98]. Esta función de activación introduce un parámetro de pendiente para los valores negativos de x. Utiliza una curva exponencial para definir los valores negativos, como se muestra en (3.3),

$$\text{eLU(x)} = \begin{cases} x & \text{if } x \geq 0 \\ a(\exp(x) - 1) & \text{if } x < 0 \end{cases} . \tag{3.3}$$

La función exponencial ayuda a suavizar los valores negativos y evitar que las neuronas se vuelvan inactivas. Se ha demostrado que ELU mejora el rendimiento en ciertos tipos de modelos de aprendizaje profundo en comparación con ReLU [97, 98].

PReLU (Unidad Lineal Rectificada Paramétrica): PReLU es una variación de la función de activación ReLU que introduce un parámetro aprendible α , que controla la pendiente de la parte negativa de la función [97, 98], como se muestra en (3.4),

$$\text{PReLU(x)} = \begin{cases} x & \text{if } x \geq 0 \\ ax & \text{if } x < 0 \end{cases} \quad . \tag{3.4}$$

PReLU puede verse como una generalización de ReLU, donde el parámetro α permite que la función aprenda una pendiente diferente para cada neurona. PReLU ha demostrado mejorar el rendimiento de las redes neuronales profundas en ciertas tareas, especialmente cuando los datos son escasos o ruidosos [97, 98].

Con el objetivo de realizar entrenamientos más cortos y descartar individuos que no obtengan un buen desempeño, se seleccionó un conjunto reducido de imágenes del conjunto de datos VGGFace2 para entrenar cada arquitectura de las CNNs para un rango de orientación de pose dentro del Algoritmo Genético. El número de imágenes seleccionadas para cada rango de orientación de pose fue el siguiente: Algoritmo genético para pequeñas rotaciones: 392.364 imágenes; Algoritmo genético para rotaciones medianas: 401.243 imágenes; Algoritmo genético para rotaciones grandes: 122.527 imágenes. La cantidad de imágenes utilizadas para el entrenamiento del algortimo genético para las rotaciones grandes fue menor porque el conjunto de datos tiene menos imágenes para ese rango de orientación.

Para el entrenamiento de cada AG, se utilizó una GPU modelo RTX 2080 Ti. Este hardware específico fue seleccionado por su capacidad para manejar de manera eficiente las complejas operaciones computacionales requeridas en nuestro estudio. Se observa que el proceso de entrenamiento consumió aproximadamente dos días por generación, específicamente en los casos de rotaciones pequeñas y medianas. Esto refleja la intensidad computacional requerida para ejecurar cada AG.

Se utilizó el método img2pose [59][99] para estimar las poses en el conjunto de datos VGGFace2. img2pose es un método desarrollado para estimar la pose de rostros en imágenes utilizando redes neuronales convolucionales. La pose de un rostro se refiere a su orientación en el espacio tridimensional, es decir, su ángulo de inclinación y dirección de la mirada. Estimar la pose de un rostro es importante en aplicaciones de reconocimiento facial, realidad virtual, y otras aplicaciones relacionadas con la percepción espacial.

El método img2pose utiliza una red neuronal convolucional que procesa la imagen del rostro para estimar su pose en seis grados de libertad: rotación en los ejes x, y, z, y traslación en los ejes x, y, z. La red es entrenada en una base de datos de imágenes de rostros con pose conocida y utiliza técnicas de regresión para estimar la pose de los rostros en imágenes nuevas que no han sido utilizadas en los entrenamientos.

El método img2pose ha demostrado ser altamente efectivo en la detección de pose de rostros en imágenes, superando a otros métodos existentes en términos de error de estimación de pose y velocidad de procesamiento en las principales bases de datos, como AFLW-2000 3D y BIWI. Además, es capaz de detectar la pose de múltiples rostros en una sola imagen, lo que lo hace útil en aplicaciones que involucran la detección y seguimiento de múltiples personas. Se propone un enfoque novedoso que estima la pose facial en 6 grados de libertad (6DoF) en 3D para todos los rostros presentes en una imagen de manera directa y sin la necesidad de una detección previa de rostros. Se implementa un método eficiente de conversión de pose para mantener la consistencia entre las estimaciones y poses de referencia entre una imagen y sus propuestas ad-hoc, y se demuestra cómo las estimaciones generadas de pose en 3D pueden convertirse en cajas delimitadoras precisas en 2D como subproducto, con una carga computacional mínima.

La figura 3.1 muestra algunos resultados cualitativos de estimación de pose sobre imágenes arbitrarias de la base de datos WIDER FACE [99].

Figura 3.1: Ejemplos de estimación de pose del rostro para imágenes del set de validación de la base de datos WIDER FACE utilizando el método img2pose. Imagen extraída de [59].

Se utilizó este enfoque para estimar la pose en el conjunto reducido de imágenes, en cada rango de orientación de pose para entrenar cada GA. El mismo método img2pose se utilizó

para estimar la pose para sesiones de entrenamiento más largas de las mejores arquitecturas encontradas en ambas particiones de entrenamiento de VGGFace2 y MS1M. Además, se utilizó img2pose durante las pruebas en la partición de prueba del conjunto de datos VGGFace2.

Se evaluó el rendimiento de cada arquitectura de red neuronal convolucional (CNN) en el conjunto de pruebas del conjunto de datos VGGFace2, calculando el porcentaje de aciertos como la proporción de imágenes reconocidas correctamente frente al número total de imágenes de prueba. También se midió el desempeño de FR en los siguientes conjuntos de datos, ampliamente utilizados con los métodos SOTA, incluyendo, VGGFace2_FP, Cross-Pose LFW (CPLFW) [60] [2] y Celebrities Frontal-Profile Faces (CFP) [61][3]. Se siguieron los protocolos de evaluación establecidos para cada conjunto de datos. La cantidad de imágenes de cada conjunto se muestra en la Tabla 3.2.

Las pruebas de FR se puede categorizar en verificación e identificación facial. En ambos escenarios, un conjunto de sujetos conocidos se enrola inicialmente en el sistema, llamando a este grupo galería, y durante las pruebas, se presenta un nuevo rostro que no forma parte de la galería. Después de que las redes profundas son entrenadas con las bases de datos de entrenaiento bajo la supervisión de una función de pérdida adecuada, cada una de las imágenes de prueba sirve como dato de entrada a las redes para obtener una representación de características. Utilizando la distancia coseno o la distancia L2, la verificación de rostros calcula la similitud uno a uno entre la galería y el rostro de prueba para determinar si las dos imágenes corresponden o no al mismo sujeto, mientras que la identificación de rostro calcula la similitud uno a muchos para determinar la identidad específica de un rostro de prueba [6, 36, 35].

Se realizaron dos conjuntos diferentes de experimentos para evaluar el rendimiento de FR con el método propuesto, utilizando las tres CNNs estructuradas, cada una con una GA para un rango de orientación facial particular. En el primer conjunto de experimentos, se probó el desempeño de FR para comparar nuestros resultados con los del SOTA utilizando la partición de prueba estándar de VGGFace2 con varias configuraciones de entrenamiento, utilizando los conjuntos de datos estándar VGGFace2 y MS1M para el entrenamiento. El segundo conjunto de experimentos evaluó el rendimiento del porcentaje de aciertos de FR de las tres CNNs estructuradas con las GAs, utilizando los dos rangos de orientación que no se usaron para desarrollar la estructura de la CNN. Por lo tanto, estas pruebas tienen como objetivo mostrar si cada CNN funciona mejor en desempeño de FR en el conjunto de prueba que contiene el mismo rango de orientación utilizado para generar la estructura de la CNN.

Tabla 3.2: Cantidad de imágenes disponibles en los conjuntos de datos ampliamente utilizados para medir el desempeño de reconocimiento de rostros por los métodos del estado del arte.

Dataset	Usage	# IDs	# Images	Key Features
VGGFace2 [81]	Train	8.6k	3.1M	depth; head part of long tail; cross pose, age and ethnicity; celebrity
MS1MV2 [16]	Train	85k	5.8M	Breadth, central part of long tail, celebrity
VGGFace2 [81]	Val	500	173k	depth; head part of long tail; cross pose, age and ethnicity; celebrity
VGGFace2 [81]	Test	500	5,000 pairs	Cross-pose
CPLFW [85]	Test	5,749	6,000 pairs	Cross-pose
CFP_FP [86]	Test	500	7,000 pairs	Frontal-profile

3.1. Experimentos

3.1.1. Experimento 1

En el Experimento 1, luego de seleccionar las 3 CNNs que fueron estructuradas a través de cada algoritmo genético, cada uno para FR dentro de un rango de orientación de pose facial particular, se entrenaron con el conjunto de datos VGGFace2 (partición de entrenamiento) y se evaluaron con el conjunto de datos VGGFace2 (partición de evaluación). También se entrenaron los tres mejores CNNs que resultaron de los GAs con el dataset MS1MV2 y se evaluaron en los diferentes conjuntos de datos indicados en la Tabla 3.2 (VGGFace2_FP, CPLFW y CFP). Este experimento nos permitió determinar la efectividad de nuestro enfoque en el porcentaje de aciertos de FR y compararla con los resultados publicados en el SOTA. La Figura 3.2(a) muestra un diagrama que ilustra la evaluación de los diferentes CNNs en el Experimento 1. En este experimento, se entrenaron tres CNNs separados en conjuntos de datos reducidos que contenían ejemplos dentro de rangos de orientación de pose facial específicos. Se seleccionaron los mejores individuos de cada generación en el GA y se entrenaron por más tiempo utilizando la partición de datos de entrenamiento VGGFace2. La Figura 3.2 muestra que se realizó entrenamiento adicional utilizando el conjunto de datos de entrenamiento MS1MV2 y se logró la evaluación en los conjuntos de datos indicados en la Tabla 3.2 (VGGFace2, CPLFW y CFP).

El componente 'Enrolled DB' del sistema está diseñado para almacenar los embeddings generados para cada individuo enrolado, correspondientes a las tres categorías distintas de

poses faciales. En el proceso de reconocimiento, cuando se recibe una imagen de entrada, el primer paso consiste en la estimación del ángulo del rostro presente en la imagen. Posteriormente, se calcula el embedding utilizando la CNN adecuada, seleccionada en función del rango de rotación facial estimado (pequeño, mediano o grande). Finalmente, este embedding recién calculado se compara con los embeddings correspondientes previamente almacenados en 'Enrolled DB'. Este método asegura que la comparación se realice con los embeddings más relevantes y específicos para la pose estimada en cada imagen de prueba.

3.1.2. Experimento 2

En el Experimento 2, se evaluó el desempeño de cada una de las 3 CNNs diseñadas para un rango específico de orientación de rostros. Se evaluó el desempeño de reconocimiento facial de la CNN diseñada para el rango de orientación de rostros pequeño utilizando las particiones de prueba de los rangos de orientación de rostros medio y grande de la base de datos VGGFace2. De manera similar, se evaluó el porcentaje de aciertos del reconocimiento facial de la CNN diseñada para el rango de orientación de rostros medio utilizando las particiones de prueba de los rangos de orientación de rostros pequeño y grande. Finalmente, se midió el desempeño de FR de la CNN diseñada para el rango de orientación de rostros grande utilizando las particiones de prueba de los rangos de orientación de rostros pequeño y medio. El objetivo principal del Experimento 2 fue verificar si cada CNN diseñada se desempeña mejor en el rango de orientación de rostros para el cual fue diseñada específicamente. Al evaluar el desempeño del reconocimiento facial de cada CNN en las particiones de prueba de diferentes rangos de orientación de rostros en la base de datos VGGFace2, se pretende demostrar que una CNN entrenada en un rango específico de orientación de rostros es más efectiva al probarla en rostros dentro del mismo rango.

Complementando lo anterior, se analizaron las arquitecturas resultantes de las CNN y las diferencias observadas en los mejores individuos obtenidos a lo largo del proceso evolutivo. El enfoque estuvo específicamente en examinar los cambios en los principales parámetros y las diferencias que definen las mejores arquitecturas de las CNN obtenidas en cada GA, como el número de conexiones, el número de céldas y bloques, comparando los hiperparámetros entre las arquitecturas de las CNNs de las generaciones iniciales con las obtenidas en las últimas generaciones. Además, se investigaron otros factores que pueden diferenciar las arquitecturas en el proceso evolutivo, como el número de convoluciones con diferentes tamaños de kernel, el número de capas de reducción y los tipos de conexiones más comunes en cada célula. A través de este análisis, se identificaron algunos de los factores que contribuyeron al éxito en la evolución de las arquitecturas de las CNNs y al mejoramiento de la precisión de FR en

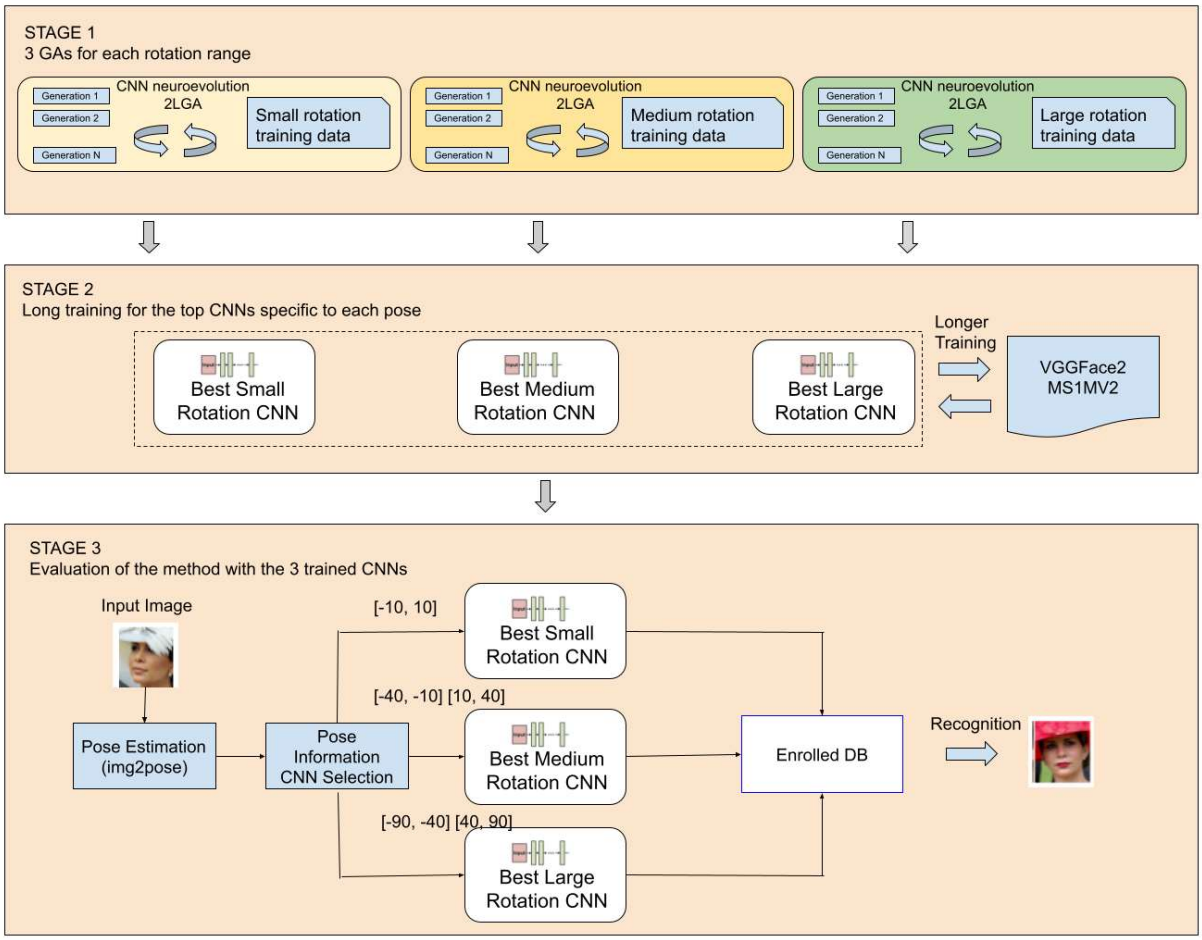


Figura 3.2: Diagrama de entrenamiento para cada Algoritmo Genético (GA) aplicado a un rango específico de orientación de pose facial. El método propuesto consta de tres etapas: En la Etapa 1, se emplean tres GAs diferentes para obtener tres CNNs separadas, cada una diseñada para manejar un rango particular de orientaciones de pose facial. En la Etapa 2, cada una de las redes obtenidas se entrena ampliamente utilizando el conjunto de entrenamiento VGGFace2 o el conjunto de datos MS1MV2, dependiendo del experimento específico. En la Etapa 3, se evalúa el rendimiento del método utilizando la partición de pruebas de VGGFace2, con cada una de las tres mejores CNN evaluadas en un rango específico de orientaciones de pose facial. Para las pruebas, se utilizan imágenes de los conjuntos de datos VGGFace2_FP, CPLFW y CFP en diferentes experimentos.

diversos cambios de pose.

También se investigarán las principales diferencias entre las mejores CNN obtenidas de los tres GAs, específicas para un rango particular de orientación de pose facial. Se cuantificaron las diferencias entre los parámetros de la red para resaltar aquellos cambios que produjeron una mejora de desempeño en un rango específico de orientación de pose facial. Se cuantificaron

las diferencias entre las arquitecturas para cada una de las tres CNN con el objetivo de identificar qué parámetros de CNNs eran relevantes para cada rango de orientación de pose.

3.2. Bases de datos

Dos de los conjuntos de datos más utilizados para entrenamiento en la comunidad científica son VGGFace2 y MS-Celeb-1M [4, 62]. VGGFace2 [4] consta de 3,31 millones de imágenes de rostros de 9131 individuos. El conjunto de datos VGGFace2 se divide en dos partes, donde el conjunto de entrenamiento tiene 8631 individuos mientras que el conjunto de evaluación tiene solo 500 individuos. MS-Celeb-1M [62] originalmente contenía una gran cantidad de ruido, pero una versión refinada se hizo pública y está disponible en [96]. Este conjunto de datos incluye alrededor de 85.000 sujetos con 5,8 millones de imágenes alineadas y comúnmente se utiliza en investigaciones debido a su gran tamaño y alta calidad de las imágenes.

Los conjuntos de datos LFW [18], CPLFW [2] y CFP [3] son ampliamente utilizados en la investigación de reconocimiento de rostros para fines de prueba. Por lo tanto, se emplearon estos conjuntos de datos para comparar nuestros resultados con los del SOTA.

El dataset Cross-Pose LFW (CPLFW) [2] se deriva de la base de datos LFW con variación intraclase de pose para las 3000 parejas de rostros positivos, mientras se asegura que las otras 3000 parejas de rostros negativos tengan los mismos atributos de género y raza.

La figura 3.3 muestra ejemplos de pares positivos de la base de datos CPLFW, que es una de las utilizadas para medir el desempeño del método. Se puede apreciar que en CPLFW la variabilidad de poses es mucho mayor que en los pares utilizados en la base de datos LFW.

Esto se puede ver de forma más clara en la figura 3.4 al visualizar la distribución de poses del conjunto CPLFW (azul), en comparación al dataset LFW (rojo), lo que lo convierte en un conjunto ideal para probar métodos que estén orientados a reconocimiento de rostros con robustez a poses extremas.

El conjunto de datos CFP [3] contiene 7000 imágenes de rostros de 500 identidades distintas, donde cada identidad tiene 10 imágenes de rostros frontales y 4 imágenes de perfil. Hay 3500 parejas de rostros positivos y 3500 parejas de rostros negativos. Las evaluaciones se realizan según dos protocolos, denominados frontal-frontal (CFP-FF) y frontal-perfil (CFP-FP). La figura 3.5 muestra ejemplos de pares positivos incluídos en esta base de datos.

Para VGGFace2 [4] hay disponibles dos protocolos de evaluación, que incluyen la evaluación de las variaciones de pose y edad. La figura 3.6 muestra las principales estadísticas

de variaciones de pose y edad para esta base de datos. En este caso, se escogió la partición VGG2-FP, que enfatiza la variación de pose, y consiste en evaluar, de manera similar a las bases de datos CPLFW y CFP, 2500 pares de rostros positivos y 2500 negativos.

El protocolo de evaluación proporcionado por cada conjunto de datos se aplicará en cada caso para medir el desempeño del método implementado. Las estadísticas del conjunto de pruebas se resumen en la Tabla 3.2.

Para los experimentos realizados, se extraen embeddings de la imagen original utilizando las mejores arquitecturas de CNNs encontradas por el AG, con la finalidad de realizar la identificación y verificación facial, y se empleó el método de comparación de similitud con la distancia coseno. La verificación 1:1 y el clasificador del vecino más cercano se utilizan para la verificación e identificación facial, respectivamente.

Figura 3.3: Ejemplos de pares positivos utilizados en la evaluación de la base de datos CPLFW y su comparación con ejemplos de pares utilizados en LFW. Imagen extraída de [2].

Además, en la figura 3.4 (gráfico de la derecha) se aprecia que en la base de datos CPLFW las diferencias de pose son significativamente mayores en comparación a los pares de rostros de la base de datos LFW. En LFW, el valor máximo de diferencia entre dos imágenes de un par positivo son 40°, mientras que en CPLFW esa diferencia llega hasta los 100°.

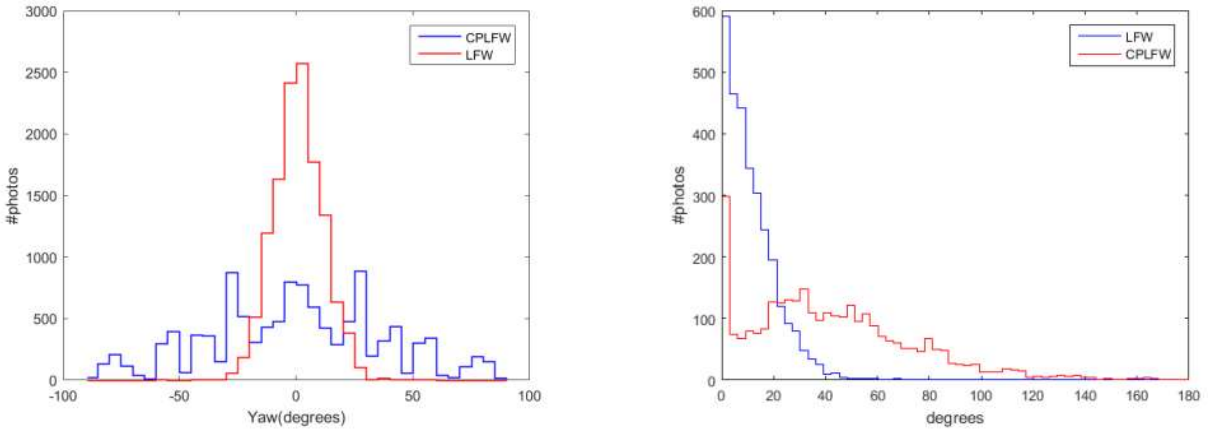
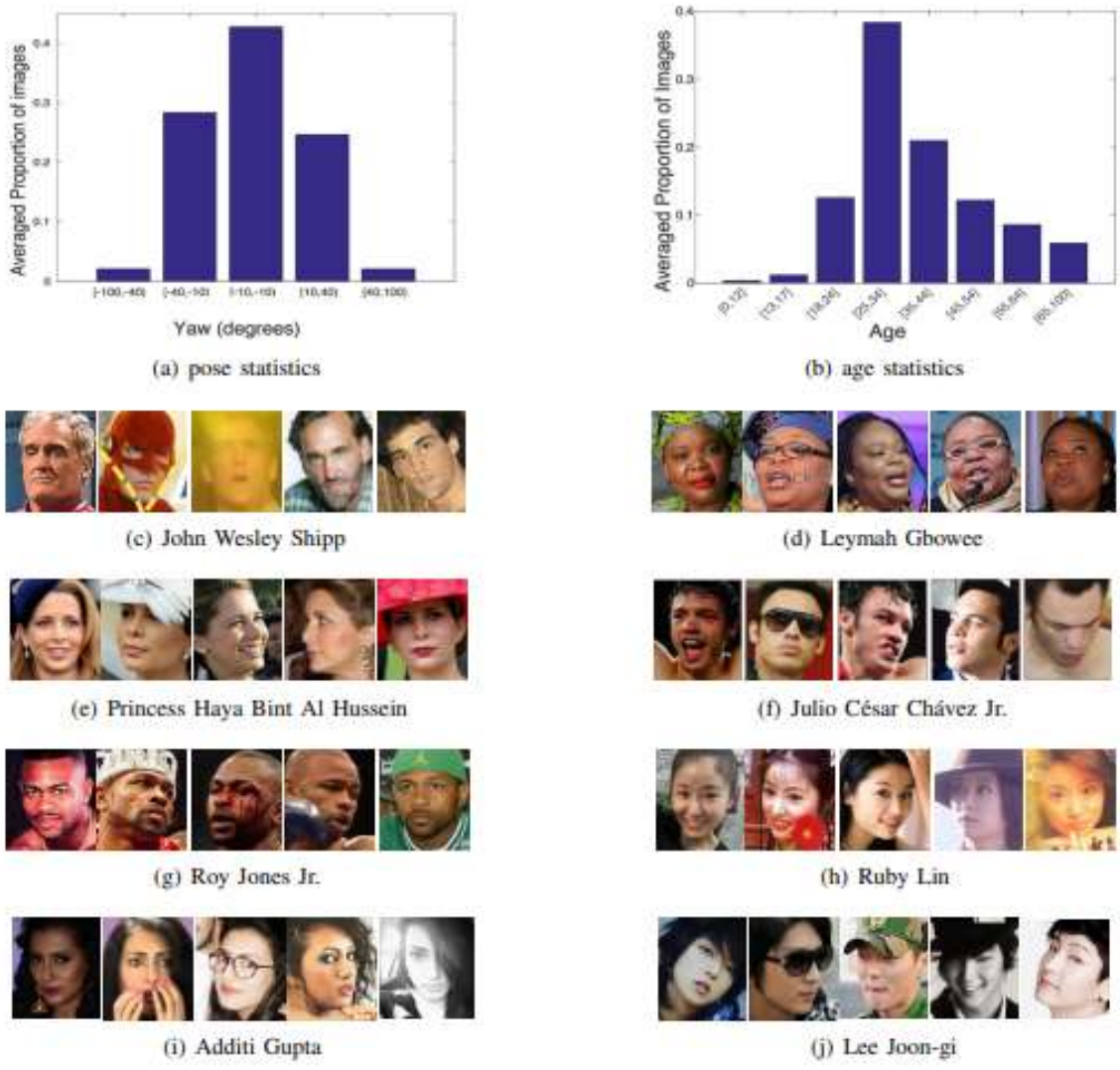


Figura 3.4: Izquierda: Comparación en la distribución de poses de la base de datos LFW (rojo) y CPLFW (azul). Derecha: Diferencia de grados en los pares positivos de la base de datos LFW (azul) y CPLFW(rojo).

Figura 3.5: Ejemplos de pares positivos incluídos en la base de datos CFP. Imagen extraída de [3].



(a) pose statistics

(b) age statistics

(c) John Wesley Shipp

(d) Leymah Gbowee

(e) Princess Haya Bint Al Hussein

(f) Julio César Chávez Jr.

(g) Roy Jones Jr.

(h) Ruby Lin

(i) Additi Gupta

(j) Lee Joon-gi

Figura 3.6: VVGFace2. a) Estadísticas de pose b) estadísticas de edad (c-j) ejemplos de 8 individuos con distintas etnias. Imagen extraída de [4].

Capítulo 4

Resultados, análisis y discusión

Cada algoritmo genético diseña la arquitectura de la CNN para un rango particular de orientación de pose facial (pequeñas rotaciones, desde -10° hasta 10°; rotaciones medianas, desde -10° hasta -40° y desde 10° hasta 40°; y grandes rotaciones, desde -40° hasta -90° y desde 40° hasta 90°) a lo largo de todas las generaciones, población inicial y valores de parámetros definidos en 3.1. La Figura 4.1 muestra los resultados de desempeño de reconocimiento de rostros de la evolución del mejor individuo, para cada generación y para cada uno de los tres algoritmos genéticos. El individuo con el mejor desempeño en cada generación de cada CNN logró el mejor desempeño de reconocimiento en una partición específica del conjunto de datos VGGFace2. Cada una de las tres particiones corresponde a un rango diferente de orientación de pose facial. La curva roja muestra el desempeño de los mejores individuos para el AG entrenado en grandes rotaciones, desde -40° hasta -90° y desde 40° hasta 90°, mientras que la curva verde representa a los mejores individuos para el AG entrenado en rotaciones medianas, desde -10° hasta -40° y desde 10° hasta 40°. La curva azul muestra a los mejores individuos para el AG entrenado en pequeñas rotaciones, desde -10° hasta 10°. Los resultados en la Figura 4.1 muestran que las CNN mejoran el desempeño de FR con la evolución de cada AG como una función de cada generación. Hay que tener en cuenta que para los tres AGs, el entrenamiento de las CNNs se realizó considerando un máximo de solo 54 épocas por individuo para reducir el tiempo computacional. Posteriormente, los mejores individuos se entrenaron usando 150 épocas para alcanzar un máximo desempeño de reconocimiento.

Después de identificar a los individuos con mejor desempeño de reconocimiento en el conjunto de VGGFace2 utilizado como validación descrito en la 3.2 para cada generación del AG, para cada uno de los tres rangos de orientación de pose facial, se lleva a cabo un proceso de entrenamiento extendido (hasta 150 épocas). Utilizamos el subconjunto correspondiente (pequeñas, medianas y grandes rotaciones) de los conjuntos usados como entrenamiento y

validación del dataset VGGFace2.

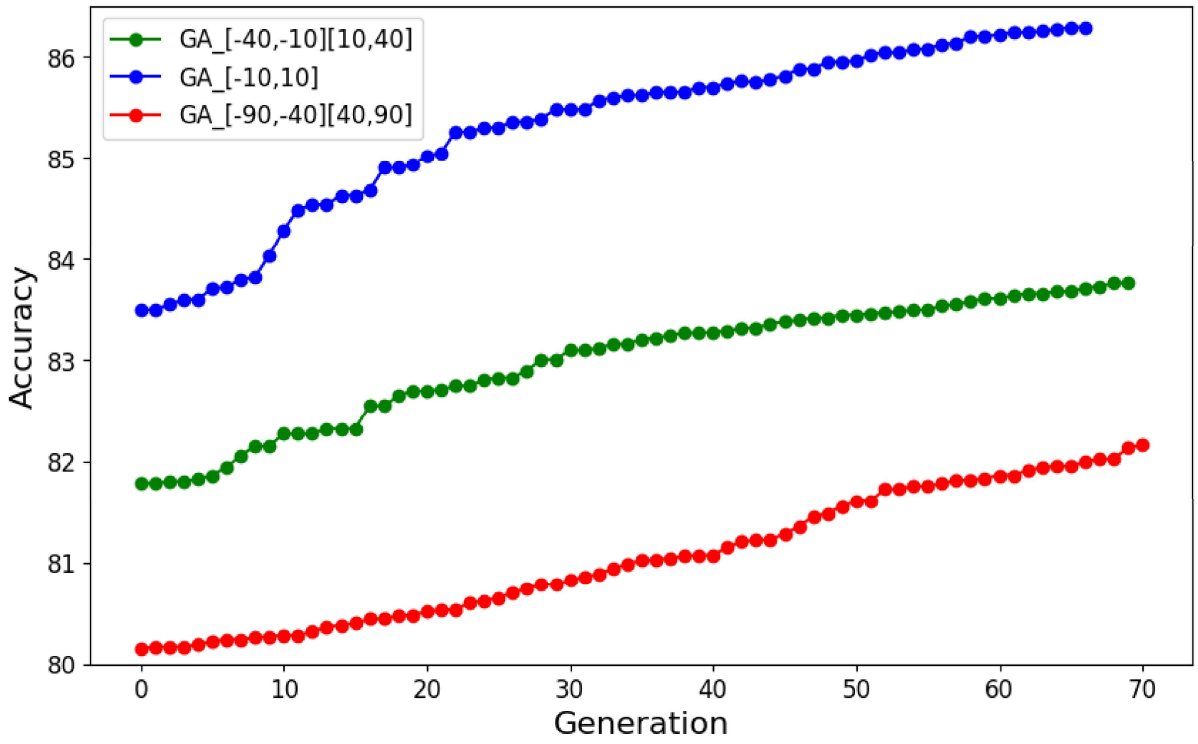


Figura 4.1: Desempeño para el mejor individuo de cada generación de cada uno de los tres AGs entrenados en una partición del conjunto de datos VGGFace2, y probados en la partición de VGGFace2 utilizada como validación y descrita en la tabla 3.2, para los tres rangos de orientación de pose facial. Rango de rotación grande (rojo), rango de rotación mediano (verde) y rango de rotación pequeño (azul). Tener en cuenta que para el AG, el entrenamiento de cada CNN incluyó un máximo de solo 54 épocas para reducir el tiempo computacional.

En la Tabla 4.1 se presentan 4 ejemplos de los resultados obtenidos a partir de la ejecución del Experimento 1 para los 3 AGs en las 3 particiones (rangos de rotación pequeño, mediano y grande) del conjunto de pruebas de VGGFace2. El mejor resultado en el conjunto de VGGFace2 utilizado como validación se alcanzó con el mejor individuo de la generación 59 para las rotaciones pequeñas, la generación 64 para las rotaciones medianas y la generación 69 para las rotaciones grandes. Siguiendo la idea de regiones del cerebro especializadas en un rango de orientación facial específico, utilizamos tres CNNs que produjeron un desempeño de reconocimiento del 95.70 % en el conjunto de datos de prueba de la base de datos VGGFace2.

Los resultados presentados en la Tabla 4.1 muestran las mejoras en el desempeño de FR de nuestro enfoque en la partición de pruebas del conjunto de datos VGGFace2 en comparación con los enfoques descritos en el SOTA. Como nuestro método utiliza tres CNNs, cada una para un rango particular de orientación de la pose facial. Este enfoque resultó en una tasa

Tabla 4.1: Resultados del Experimento 1 utilizando 3 AGs en el conjunto de pruebas VGG-Face2. Cada fila muestra los resultados obtenidos con 3 CNN estructuradas para un número diferente de generaciones para las 3 rangos de orientación de pose facial. Se muestra el desempeño de FR para cada rango de orientación, y en la última columna se muestra el desempeño de FR total para la prueba VGGFace2.

Method	**[-90 ,-40]**	**[-40, -10]**	**[-10, 10]**	**[10, 40]**	**[40, 90]**	**Acc.**
2LGA_FaceRec 1	89.41	94.21	98.74	93.92	90.02	95.64
2LGA_FaceRec 2	89.83	94.31	98.43	94.51	90.17	95.69
2LGA_FaceRec 3	89.61	94.29	98.8	93.97	90.22	95.70
2LGA_FaceRec 4	90.08	94.44	98.76	94.65	90.38	95.91

de reconocimiento de 95.91 % en el conjunto de VGGFace2 utilizado para validar.

Además, la Tabla 4.2 muestra los resultados de nuestro método utilizando los conjuntos de verificación CPLFW, CFP_FP y VGGFace2_FP, y empleando el conjunto de entrenamiento MS1MV2. Todos los métodos expuestos en esta tabla, incluyendo los métodos propuestos, fueron entrenados con el dataset MS1MV2. Los resultados muestran que se lograron mejoras en comparación con las publicaciones SOTA (State-of-the-Art). En el conjunto de prueba CPLFW, el error de verificación se redujo en un 0.67 %, lo que representa una mejora del 11.04 % (% de error). En el conjunto de prueba CFP_FP, el error de verificación se redujo en un 0.07 %, logrando una mejora del 7.87 % (% de error). Además, en el conjunto de prueba VGGFace2_FP, el error de verificación se redujo en un 0.18 %, lo que representa una mejora del 4.04 % (% de error).

Por ejemplo, la mejora alcanzada en el conjunto de prueba CPLFW, entre SOTA [44] y [82], fue del 6.9 % (0.45/6.52) y, por lo tanto, nuestra mejora es mayor que la reportada anteriormente. De manera similar, la mejora alcanzada por SOTA [85] en comparación con [44] fue del 8.2 % (0.08/0.97) en CFP_FP y del 1.1 % (0.05/4.5) en VGGFace2_FP.

En el estudio de ablación, evaluamos nuestro método midiendo el rendimiento al utilizar la misma CNN para cada uno de los tres rangos de orientación de pose facial. Por lo tanto, una de las tres CNN se utilizó de manera individual para el reconocimiento/verificación. Se realizó una comparación de rendimiento con todos los módulos. Este estudio de ablación sirvió para respaldar los resultados presentados en la Tabla 4.3, que indica que cada CNN específica está adaptada para un rango de pose particular. El estudio de ablación confirmó la disminución significativa en el rendimiento al aplicar cada red de manera individual a los conjuntos de datos de verificación. En el conjunto de datos CPLFW, nuestro método,

Tabla 4.2: Comparación de rendimiento en los conjuntos de datos SOTA, CFP_FP, CPLFW y VGGFace2_FP, utilizando como dataset de entrenamiento MS1MV2

Method	**CFP_FP**	**CPLFW**	**VGGFace2_FP**
IR152_Arcface [96]	98.37	93.05	95.50
IR152_Adacos [100]	98.42	92.57	95.43
IR152_AM-Soft. [101]	98.45	92.52	95.55
LighCNN-29v2_AF [96]	94.35	85.78	94.06
ResNet50_AF [96]	98.27	93.48	-
AFRN [102]	95.56	93.48	-
BroadFace [44]	98.63	93.38	-
MagFace [45]	98.46	92.87	-
VPL-Arcface [46]	99.11	93.45	-
AdaFace [49]	99.03	93.93	-
2LGA_FaceRec 2	99.1	94.45	95.60
2LGA_FaceRec 3	99.11	94.47	95.62
2LGA_FaceRec 4	99.18	94.60	95.73

Tabla 4.3: Estudio de ablación en los conjuntos de datos CFP_FP, CPLFW y VGGFace2_FP utilizando la base de datos MS1M para entrenamiento.

Method	CFP_FP	CPLFW	VGGFace2_FP
2LGA_FaceRec 4	99.18	94.60	95.73
Small Rotation CNN	95.44	91.01	92.32
Medium Rotation CNN	95.86	91.59	93.36
Large Rotation CNN	90.53	83.63	84.11

que incorpora las tres redes y la información de pose, alcanza un rendimiento del 94.60 %. Sin embargo, al utilizar una sola red para la tarea de verificación, hubo una disminución a 91.01 %, 91.59 % y 83.63 % para la CNN de rotación pequeña, la CNN de rotación mediana y la CNN de rotación grande, respectivamente. De manera similar, en el conjunto de datos de prueba CFP_FP, nuestro método alcanza un rendimiento del 99.18 %. Sin embargo, al utilizar una sola red para la tarea de verificación, hubo una disminución a 95.44 % para la CNN de rotación pequeña, 95.86 % para la CNN de rotación mediana y 90.53 % para la CNN de rotación grande. Con el conjunto de datos de prueba VGGFace2_FP, nuestro método alcanza un rendimiento del 95.73 %. Sin embargo, al utilizar una sola red para la tarea de verificación, se observó una disminución a 92.32 %, 93.36 % y 84.11 % para la CNN de rotación pequeña, la CNN de rotación mediana y la CNN de rotación grande, respectivamente. Como

se muestra en la Tabla 4.3, la integración de las tres CNN entrenadas para cada rango de pose particular mejora significativamente el rendimiento en los conjuntos de datos de prueba VGGFace2_FP, CFP_FP y CPLFW, en comparación con el uso de solo una CNN para la tarea de verificación facial. Además, es significativo que el uso de individuos de generaciones tempranas, especialmente a partir de la generación 3, alcanza un rendimiento mucho menor en comparación con los resultados utilizando generaciones posteriores. Esto confirma que el AG es capaz de encontrar arquitecturas optimizadas para cada rango de pose.

En el Experimento 2, se midió el desempeño de reconocimiento facial de cada uno de los 3 mejores individuos hallados con los 3 distintos AGs para un rango de orientación de pose facial diferente al utilizado para cada evolución. Este experimento verifica si cada arquitectura de CNN alcanzó los mejores resultados para el rango específico para el cual se desarrolló y entrenó. El experimento se realizó utilizando el conjunto de entrenamiento y test de la base de datos VGGFace2. Los resultados en la Tabla 4.4 muestran que las CNNs tienen un rendimiento superior para el rango de orientación de pose facial específico en el que cada una fue generada y entrenada. Por ejemplo, la CNN obtenida para el rango de rotación mediano exhibe un desempeño de reconocimiento superior para ese rango de orientación de pose, en comparación con los rangos de rotación pequeño y grande. De manera similar, la CNN desarrollada para el rango de rotación pequeño muestra un mejor rendimiento para el rango de rotación pequeño, en comparación con los rangos moderado y grande. Este resultado confirma la eficacia de nuestro enfoque en el diseño, a través de la evolución, de CNN que tienen resultados mejorados para rangos específicos de orientación de pose.

Tabla 4.4: Los resultados de la validación de poses cruzadas, Experimento 2, en el conjunto de datos VGGFace2 usando uno de nuestros métodos (2LGA_FaceRec 3).

CNN	**[-90 ,-40]**	**[-40, -10]**	**[-10, 10]**	**[10, 40]**	**[40, 90]**	**Acc.**
Small	80.92	89.91	98.8	89.26	80.96	92.97
Medium	83.27	94.29	94.09	93.97	83.31	93.43
Large	89.61	85.07	83.89	84.74	90.22	84.40

En la Tabla 4.5 se presentan los principales parámetros que definen las mejores arquitecturas de CNN encontradas con cada uno de los 3 algoritmos genéticos para un rango específico de orientación de pose. Se muestran el número de bloques (nblocks), número de celdas (ncells), número de nodos por celda (nodos/celda), número de conexiones por celda (nconn/cell) y número de bloques convolucionales (NCB).

Este resultado confirma que cada arquitectura obtenida posee diferencias en términos de cantidad de conexiones y bloques convolucionales, por lo que la evolución fue diferente en

cada caso, dependiendo del tipo de rango de pose con que cada entrenamiento y selección de mejores arquitecturas fue ejecutado.

Tabla 4.5: Los principales parámetros que definen las mejores arquitecturas de CNNs encontradas con cada uno de los 3 algoritmos genéticos para un rango específico de orientación de la pose facial. La red CNN evolucionó con estructuras diferentes dependiendo del rango de orientación de la pose utilizado para el entrenamiento.

	[-10 , 10]	**[10, 40]**	**[40, 90]**
nblocks	3	4	3
ncells	4	4	3
nodes/cell	7	6	7
nconn/cell	11	11	10
NCB	60	64	45

En la Tabla 4.6 se muestra la comparación de las arquitecturas obtenidas para las primeras 5 generaciones y las últimas 5 generaciones para los mejores individuos encontrados en el algoritmo genético del rango de rotación mediano (entre 10 y 40 grados). Los parámetros medidos corresponden al número de bloques (nblocks), número de celdas (ncells), nodos por celda (nodos/celda), número de conexiones por celda (nconn/cell), número de bloques convolucionales (NCB), convoluciones por celda, nodos totales en la arquitectura (total_nodes) y conexiones totales entre nodos en la arquitectura (total_conns).

En general, las arquitecturas de las primeras generaciones difieren significativamente en términos del número de convoluciones y conexiones en comparación con las últimas generaciones analizadas. El número promedio de conexiones en las primeras generaciones es de 80.4, mientras que en las últimas generaciones es de 138, lo que indica un aumento del 71 % en este parámetro. El número total de conexiones para los mejores 2 individuos de las primeras 5 generaciones varía entre 48 y 132, mientras que en los mejores 2 individuos de las últimas 5 generaciones el número total de conexiones varía entre 108 y 176. De manera similar, el número promedio de convoluciones totales en las primeras generaciones es de 39.1, mientras que en las últimas generaciones es de 60.4, lo que indica un aumento del 54.5 % en este parámetro.

Para una mejor visualización, la Figura 4.2 muestra gráficamente la diferencia en los valores de los principales parámetros de las 10 redes analizadas entre las primeras 5 generaciones y las 10 CNN obtenidas en las últimas 5 generaciones (ver tabla 4.6). Los parámetros incluyen nodos por celda, conexiones por celda, convoluciones por celda, convoluciones totales y conexiones totales. Cada parámetro se normalizó entre 0 y 1 utilizando los valores máximos

Tabla 4.6: Comparación de algunos parámetros de arquitectura obtenidos para las primeras 5 generaciones y las últimas 5 generaciones de 2 arquitecturas CNN: (a) número de bloques, (b) número de células, (c) nodos por célula, (d) número de conexiones por célula, (e) número de bloques convolucionales, (f) convoluciones por célula, (g) total de nodos en la arquitectura y (h) total de conexiones entre nodos en la arquitectura.

Generation	**a**	**b**	**c**	**d**	**e**	**f**	**g**	**h**
1	3	3	6	10	45	5	54	90
1	3	2	5	8	24	4	30	48
2	3	3	6	9	36	4	54	81
2	3	2	6	9	30	5	36	54
3	3	2	6	8	30	5	36	48
3	4	3	7	11	48	4	84	132
4	3	3	5	9	45	5	45	81
4	4	2	7	10	40	5	56	80
5	4	3	7	8	48	4	84	96
5	4	3	7	10	48	4	84	120
60	4	3	6	10	60	5	72	120
60	4	3	6	9	60	5	72	108
61	3	4	7	11	60	5	84	132
61	4	4	7	9	64	4	112	144
62	4	3	6	11	60	5	72	132
62	4	3	6	10	60	5	72	120
63	4	3	8	11	60	5	96	132
63	4	4	7	11	64	4	112	176
64	4	4	8	10	64	4	128	160
64	4	4	8	11	64	4	128	176

y mínimos obtenidos entre las generaciones 1-5 y 60-64.

Otra manera de visualizar las diferencias entre cada una de las arquitecturas obtenidas consiste en visualizar las regiones del rostro que entregan mayor información en la etapa de extracción de características. Score-CAM [5] es una técnica utilizada en la visualización de CNNs para resaltar y localizar las regiones de activación en una imagen que contribuyen significativamente a la clasificación realizada por la red. El objetivo principal del método es entender qué áreas de la imagen son relevantes para la decisión de la red sobre una clase específica. En la Figura 4.3 se muestran los resultados del método Score-CAM [5]

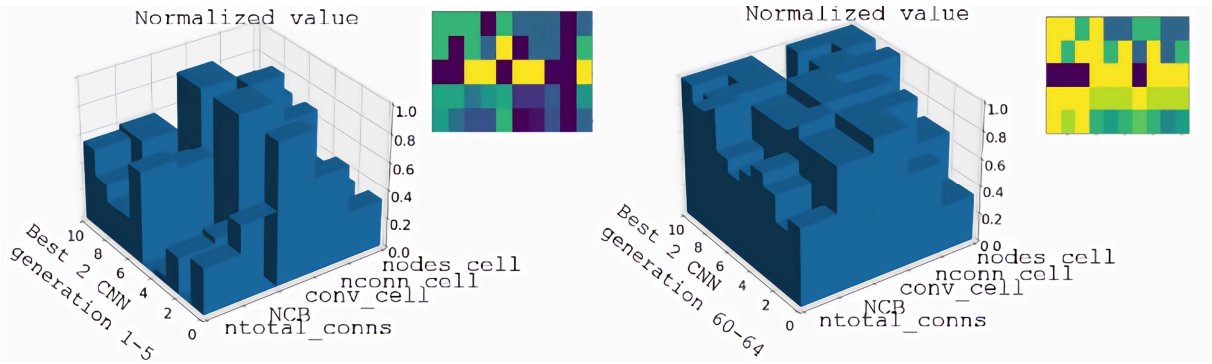


Figura 4.2: Comparación de los valores de los principales parámetros de las CNN entre las generaciones tempranas (izquierda) y las últimas (derecha) para el caso CNN rotaciones medias [10 40].

para visualizar las regiones de la imagen del rostro que son enfatizadas por cada modelo de reconocimiento facial (3 CNNs), utilizando 3 pares de rostros de la base de datos CPLFW. Se compara la visualización de la CNN de pequeñas rotaciones (+-10°) con la de la CNN de gran rotación en una cara frontal (entre 40°y 90°), y también en el caso de rostros con gran variación de pose. Como se puede observar en la Figura 4.3, las CNNs tienden a enfatizar diferentes regiones del rostro dependiendo del rango de orientación de pose, lo que proporciona algunas ideas sobre las características aprendidas por cada modelo. En otras palabras, esta visualización nos permite entender qué partes del rostro son más relevantes para cada modelo y cómo cada modelo enfatiza diferentes características faciales dependiendo de la orientación de la pose del rostro. Esta información es valiosa para comprender cómo los modelos de reconocimiento facial "ven" y analizan los rostros, y para mejorar su desempeño en diferentes situaciones y orientaciones de pose.

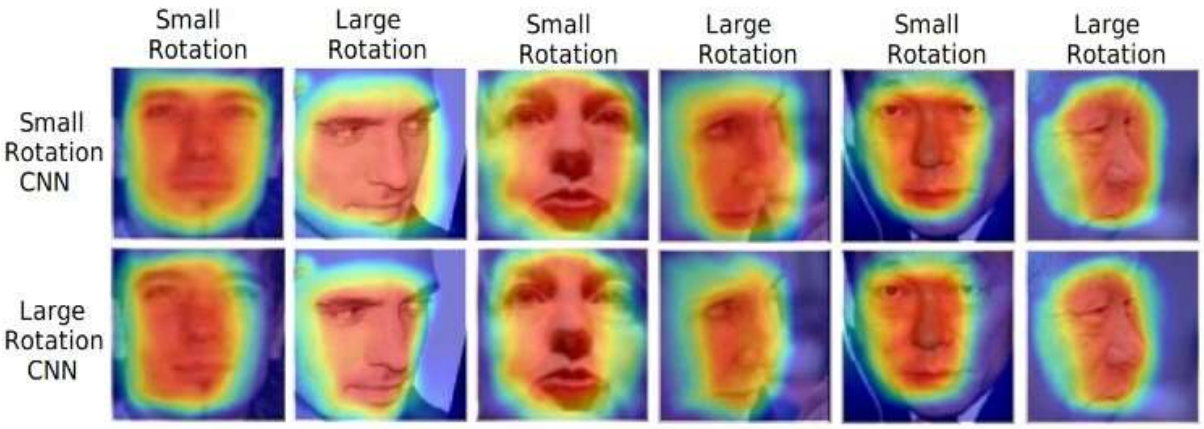


Figura 4.3: Visualización con el método Score-CAM [5] para las redes de pequeñas y grandes rotaciones en 3 pares de imágenes positivas de la base de datos CPLFW.

Capítulo 5

Conclusiones

Siguiendo la idea de las regiones especializadas en poses faciales encontradas en el cerebro, se utilizó neuroevolución empleando AGs para estructurar tres CNNs que están especializadas en un rango particular de orientación de pose facial (rotaciones pequeñas, medianas y grandes). El rendimiento de nuestro método propuesto fue evaluado en varios conjuntos de datos, cuyo uso fue reportado en publicaciones del SOTA, en FR con variaciones de pose: VGGFace2, CPFLW y CFP_FP. Estos conjuntos de datos contienen una cantidad significativa de rostros con gran variación de pose. Los GAs generaron nuevas CNN especializadas ajustadas a tres rangos de orientación facial. Los mejores individuos obtenidos en cada generación de los AGs fueron evaluados en relación a su desempeño de reconocimiento en la partición de prueba del conjunto de datos VGGFace2.

La Figura 4.1 muestra que el desempeño de reconocimiento de las tres estructuras de CNN generadas por los AG mejoró en función del número de generaciones. Los mejores individuos de las últimas generaciones fueron entrenados durante un número mayor de épocas en el conjunto de datos de entrenamiento VGGFace2. Los resultados en el conjunto de prueba del conjunto de datos VGGFace2, después de este entrenamiento prolongado, alcanzaron un desempeño de reconocimiento más alto que la publicada anteriormente en el SOTA. Las mismas mejores estructuras de CNN de la Figura 4.1 también se entrenaron utilizando el conjunto de datos MS1MV2; nuestros resultados se presentan en la Tabla 4.2 para los conjuntos de verificación VGGFace2_FP, CPLFW y CFP_FP. Los resultados alcanzaron un desempeño mayor que la del SOTA.

Los resultados muestran que las nuevas CNNs generadas con AGs mejoran el desempeño de reconocimiento en comparación con las del SOTA. Por lo tanto, la neuroevolución a través de la estrategia de 2LGA es una alternativa para diseñar nuevas CNNs que posean un buen

desempeño de reconocimiento. Además, combinar varias CNNs con la información de rotación obtenida con el módulo de estimación automática de pose, para abordar el problema de la orientación del rostro para rotaciones grandes, simulando lo que se ha observado en el cerebro con áreas especializadas en poses del rostro, produjo una mejora significativa respecto a otros trabajos del SOTA, donde la mayoría de las veces se utilizan arquitecturas *mainstream*, es decir, estructuras de CNN que son comúnmente utilizadas en tareas de FR. El porcentaje de reconocimiento fue del 95,91 % en el conjunto de pruebas del conjunto de datos VGGFace2, y del 95.73 %, 94,60 % y el 99,18 % obtenidos para los conjuntos de datos VGGFace2_FP, CPLFW y CFP_FP después del entrenamiento con MS1MV2.

El estudio de ablación confirmó la importancia de las tres CNNs diseñadas por el GA, una para cada pose, en lograr un alto desempeño en el reconocimiento/verificación facial. Estas CNNs manejan de manera efectiva las variaciones en la orientación de la pose, especialmente en comparación con el uso de la misma CNN para todas las poses faciales.

El Experimento 2 se llevó a cabo para verificar si el desempeño de reconocimiento de cada red neuronal convolucional específica, diseñada para un rango particular de orientaciones faciales, tenía resultados similares para otros rangos de orientación de rostro diferentes. Nuestros resultados muestran que todas las CNNs exhiben un desempeño superior para el rango de orientación facial específico para el cual fueron evolucionadas y entrenadas, lo que confirma la efectividad de nuestro enfoque para mejorar la precisión de reconocimiento en conjuntos de datos con gran variación de pose.

El análisis de las arquitecturas de CNNs para cada uno de los tres rangos de orientación facial muestra que estas arquitecturas difieren en complejidad y diseño, dependiendo del rango de pose facial específico. Esta información es importante para el desarrollo de sistemas de reconocimiento facial más robustos, que puedan funcionar bien en un amplio rango de variaciones de pose. Al comparar las arquitecturas desde las primeras generaciones hasta las últimas generaciones, se determinó que difieren significativamente en términos del número de conexiones y los bloques convolucionales a medida que la CNN evoluciona.

El análisis de los principales parámetros de las redes neuronales convolucionales generadas en las primeras y nuevas generaciones muestra una diferencia significativa en la arquitectura. Las nuevas generaciones tienen tanto un mayor número de conexiones como un mayor número total de convoluciones en comparación con las generaciones tempranas, con un aumento del 71 % y 54.5 %, respectivamente.

5.1. Trabajo Futuro

Como trabajo futuro se propone la posibilidad de privilegiar ciertas arquitecturas de CNNs como individuos válidos del AG, a través de una etapa previa de clasificación, que podrá ser entrenada con la información obtenida con las arquitecturas que no obtuvieron un buen desempeño de reconocimiento.

Complementando la estrategia previa, también podrán ser explorados otros rangos de valores para los individuos, escogiendo rangos más acotados o específicos dependiendo del rango de pose que desea ser resuelto, considerando que hasta ahora el espacio de búsqueda es el mismo para los 3 algoritmos genéticos.

Por otra parte,se investigarán las posibles limitaciones en el reconocimiento facial con gran variación de pose producidas por conjuntos de datos de entrenamiento desequilibrados. Esta limitación es el resultado de conjuntos de datos de entrenamiento que presentan significativamente menos muestras de rostros con gran variación de pose. Esta subrepresentación de rostros con variación de pose significativa puede generar un sesgo en los modelos entrenados. En nuestro método, la arquitectura de las CNNs se verían afectadas por este sesgo. Una posible solución a este problema es el uso de aumentación de datos para equilibrar artificialmente las muestras subrepresentadas. Hay varios estudios que han logrado resultados fotorealistas, generando rostros artificialmente con variación de pose a partir de muestras reales casi frontales [37].

Otra área de investigación para abordar este problema es el uso de estrategias para tratar el problema de la distribución de dominio con cola larga. Las clases correspondientes a rostros con rotación de pose pequeña están sobre-representadas en comparación con las rostros con gran variación de pose. Por lo tanto, es pertinente investigar y desarrollar estrategias que incorporen múltiples mecanismos para equilibrar la distribución de dominio y optimizar el espacio de características. Este tipo de estrategia ha sido explorada en el problema general de reconocimiento facial [45].

Además de los enfoques mencionados, se plantea la posibilidad de extender y aplicar la metodología desarrollada en este estudio a otros desafíos dentro del ámbito del reconocimiento facial. Entre estos, destacan la identificación de la edad, clasificación de género [103] y el manejo de imágenes con oclusiones. Esta ampliación permitirá evaluar la versatilidad y eficacia de la metodología propuesta en contextos donde las características faciales presentan variaciones significativas o están parcialmente obstruidas y donde métodos de búsqueda de arquitecturas neuronales aún no ha sido ampliamente explorado.

Bibliografía

[1] M. S. Livingstone, J. L. Vincent, M. J. Arcaro, K. Srihasam, P. F. Schade, and T. Savage, "Development of the macaque face-patch system," *Nature communications*, vol. 8, no. 1, p. 14897, 2017.

[2] T. Zheng and W. Deng, "Cross-pose lfw: A database for studying cross-pose face recognition in unconstrained environments," *Beijing University of Posts and Telecommunications, Tech. Rep*, vol. 5, no. 7, 2018.

[3] S. Sengupta, J.-C. Chen, C. Castillo, V. M. Patel, R. Chellappa, and D. W. Jacobs, "Frontal to profile face verification in the wild," in *2016 IEEE winter conference on applications of computer vision (WACV)*, pp. 1–9, IEEE, 2016.

[4] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, "Vggface2: A dataset for recognising faces across pose and age," in *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, pp. 67–74, IEEE, 2018.

[5] H. Wang, Z. Wang, M. Du, F. Yang, Z. Zhang, S. Ding, P. Mardziel, and X. Hu, "Score-cam: Score-weighted visual explanations for convolutional neural networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pp. 24–25, 2020.

[6] M. Wang and W. Deng, "Deep face recognition: A survey," *Neurocomputing*, vol. 429, pp. 215–244, 2021.

[7] S. B. Ahmed, S. F. Ali, J. Ahmad, M. Adnan, and M. M. Fraz, "On the frontiers of pose invariant face recognition: a review," *Artificial Intelligence Review*, vol. 53, pp. 2571–2634, 2020.

[8] J. Zhao, Y. Cheng, Y. Xu, L. Xiong, J. Li, F. Zhao, K. Jayashree, S. Pranata, S. Shen, J. Xing, *et al.*, "Towards pose invariant face recognition in the wild," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2207–2216, 2018.

[9] G. Guo and N. Zhang, "A survey on deep learning based face recognition," *Computer vision and image understanding*, vol. 189, p. 102805, 2019.

[10] I. Adjabi, A. Ouahabi, A. Benzaoui, and A. Taleb-Ahmed, "Past, present, and future of face recognition: A review," *Electronics*, vol. 9, no. 8, p. 1188, 2020.

[11] H.-J. Hsu and K.-T. Chen, "Droneface: an open dataset for drone research," in *Proceedings of the 8th ACM on multimedia systems conference*, pp. 187–192, 2017.

[12] I. Kalra, M. Singh, S. Nagpal, R. Singh, M. Vatsa, and P. Sujit, "Dronesurf: Benchmark dataset for drone-based face recognition," in *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*, pp. 1–7, IEEE, 2019.

[13] D. Wang, C. Otto, and A. K. Jain, "Face search at scale," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 6, pp. 1122–1136, 2016.

[14] L. Tran, X. Yin, and X. Liu, "Disentangled representation learning gan for pose-invariant face recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1415–1424, 2017.

[15] H. Li and G. Hua, "Probabilistic elastic part model: a pose-invariant representation for real-world face verification," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 4, pp. 918–930, 2017.

[16] E.-J. Cheng, K.-P. Chou, S. Rajora, B.-H. Jin, M. Tanveer, C.-T. Lin, K.-Y. Young, W.-C. Lin, and M. Prasad, "Deep sparse representation classifier for facial recognition and detection system," *Pattern Recognition Letters*, vol. 125, pp. 71–77, 2019.

[17] W. Ali, W. Tian, S. U. Din, D. Iradukunda, and A. A. Khan, "Classical and modern face recognition approaches: a complete review," *Multimedia tools and applications*, vol. 80, pp. 4825–4880, 2021.

[18] G. B. Huang, M. Mattar, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database forstudying face recognition in unconstrained environments," in *Workshop on faces in'Real-Life'Images: detection, alignment, and recognition*, 2008.

[19] I. Kemelmacher-Shlizerman, S. M. Seitz, D. Miller, and E. Brossard, "The megaface benchmark: 1 million faces for recognition at scale," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4873–4882, 2016.

[20] A. Nech and I. Kemelmacher-Shlizerman, "Level playing field for million scale face recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7044–7053, 2017.

[21] T. Zheng, W. Deng, and J. Hu, "Cross-age lfw: A database for studying cross-age face recognition in unconstrained environments," *arXiv preprint arXiv:1708.08197*, 2017.

[22] T. Zheng and W. Deng, "Cross-pose lfw: A database for studying cross-pose face recognition in unconstrained environments," *Beijing University of Posts and Telecommunications, Tech. Rep*, vol. 5, no. 7, 2018.

[23] S. Sengupta, J.-C. Chen, C. Castillo, V. M. Patel, R. Chellappa, and D. W. Jacobs, "Frontal to profile face verification in the wild," in *2016 IEEE winter conference on applications of computer vision (WACV)*, pp. 1–9, IEEE, 2016.

[24] Y.-H. Huang and H. H. Chen, "Deep face recognition for dim images," *Pattern Recognition*, vol. 126, p. 108580, 2022.

[25] Y. Akbari, N. Almaadeed, S. Al-Maadeed, and O. Elharrouss, "Applications, databases and open computer vision research from drone videos and images: a survey," *Artificial Intelligence Review*, vol. 54, pp. 3887–3938, 2021.

[26] S. Zhang, Q. Zhang, Y. Yang, X. Wei, P. Wang, B. Jiao, and Y. Zhang, "Person re-identification in aerial imagery," *IEEE Transactions on Multimedia*, vol. 23, pp. 281–291, 2020.

[27] A. Grigorev, S. Liu, Z. Tian, J. Xiong, S. Rho, and J. Feng, "Delving deeper in drone-based person re-id by employing deep decision forest and attributes fusion," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 16, no. 1s, pp. 1–15, 2020.

[28] S. Sengupta, J.-C. Chen, C. Castillo, V. M. Patel, R. Chellappa, and D. W. Jacobs, "Frontal to profile face verification in the wild," in *2016 IEEE winter conference on applications of computer vision (WACV)*, pp. 1–9, IEEE, 2016.

[29] O. Elharrouss, N. Almaadeed, S. Al-Maadeed, and F. Khelifi, "Pose-invariant face recognition with multitask cascade networks," *Neural Computing and Applications*, pp. 1–14, 2022.

[30] F. Boutros, N. Damer, F. Kirchbuchner, and A. Kuijper, "Elasticface: Elastic margin loss for deep face recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1578–1587, 2022.

[31] S. Anwarul and S. Dahiya, "A comprehensive review on face recognition methods and factors affecting facial recognition accuracy," *Proceedings of ICRIC 2019: Recent Innovations in Computing*, pp. 495–514, 2020.

[32] C. Foster, M. Zhao, T. Bolkart, M. J. Black, A. Bartels, and I. Bülthoff, "The neural coding of face and body orientation in occipitotemporal cortex," *NeuroImage*, vol. 246, p. 118783, 2022.

[33] M. J. Arcaro, T. Mautz, V. K. Berezovskii, and M. S. Livingstone, "Anatomical correlates of face patches in macaque inferotemporal cortex," *Proceedings of the National Academy of Sciences*, vol. 117, no. 51, pp. 32667–32678, 2020.

[34] J. P. Perez and C. A. Perez, "Face patches designed through neuroevolution for face recognition with large pose variation," *IEEE Access*, vol. 11, pp. 72861–72873, 2023.

[35] H. Du, H. Shi, D. Zeng, X.-P. Zhang, and T. Mei, "The elements of end-to-end deep face recognition: A survey of recent advances," *ACM Computing Surveys (CSUR)*, vol. 54, no. 10s, pp. 1–42, 2022.

[36] I. Masi, Y. Wu, T. Hassner, and P. Natarajan, "Deep face recognition: A survey," in *2018 31st SIBGRAPI conference on graphics, patterns and images (SIBGRAPI)*, pp. 471–478, IEEE, 2018.

[37] N. Zhu, Z. Yu, and C. Kou, "A new deep neural architecture search pipeline for face recognition," *IEEE Access*, vol. 8, pp. 91303–91310, 2020.

[38] X. Wang, "Teacher guided neural architecture search for face recognition," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, pp. 2817–2825, 2021.

[39] H. Li and G. Hua, "Probabilistic elastic part model: a pose-invariant representation for real-world face verification," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 4, pp. 918–930, 2017.

[40] K. Cao, Y. Rong, C. Li, X. Tang, and C. C. Loy, "Pose-robust face recognition via deep residual equivariant mapping," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5187–5196, 2018.

[41] G. Chen, Y. Shao, C. Tang, Z. Jin, and J. Zhang, "Deep transformation learning for face recognition in the unconstrained scene," *Machine Vision and Applications*, vol. 29, pp. 513–523, 2018.

[42] S. Cen, H. Luo, J. Huang, W. Shi, and X. Chen, "Pre-trained feature fusion and multidomain identification generative adversarial network for face frontalization," *IEEE Access*, vol. 10, pp. 77872–77882, 2022.

[43] P. Zhang, F. Zhao, P. Liu, and M. Li, "Efficient lightweight attention network for face recognition," *IEEE Access*, vol. 10, pp. 31740–31750, 2022.

[44] Y. Kim, W. Park, and J. Shin, "Broadface: Looking at tens of thousands of people at once for face recognition," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX*, pp. 536–552, Springer, 2020.

[45] Q. Meng, S. Zhao, Z. Huang, and F. Zhou, "Magface: A universal representation for face recognition and quality assessment," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14225–14234, 2021.

[46] J. Deng, J. Guo, J. Yang, A. Lattas, and S. Zafeiriou, "Variational prototype learning for deep face recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11906–11915, 2021.

[47] M. He, J. Zhang, S. Shan, M. Kan, and X. Chen, "Deformable face net for pose invariant face recognition," *Pattern Recognition*, vol. 100, p. 107113, 2020.

[48] H. Zhou, J. Liu, Z. Liu, Y. Liu, and X. Wang, "Rotate-and-render: Unsupervised photorealistic face rotation from single-view images," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5911–5920, 2020.

[49] M. Kim, A. K. Jain, and X. Liu, "Adaface: Quality adaptive margin for face recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18750–18759, 2022.

[50] Z. An, W. Deng, J. Hu, Y. Zhong, and Y. Zhao, "Apa: Adaptive pose alignment for pose-invariant face recognition," *IEEE Access*, vol. 7, pp. 14653–14670, 2019.

[51] J. Liu, Q. Li, M. Liu, and T. Wei, "Cp-gan: A cross-pose profile face frontalization boosting pose-invariant face recognition," *IEEE Access*, vol. 8, pp. 198659–198667, 2020.

[52] X. Luan, H. Geng, L. Liu, W. Li, Y. Zhao, and M. Ren, "Geometry structure preserving based gan for multi-pose face frontalization and recognition," *IEEE Access*, vol. 8, pp. 104676–104687, 2020.

[53] S. Guo, R. Liu, M. Wang, M. Zhang, S. Nie, S. Lina, and N. Abe, "Exploiting the tail data for long-tailed face recognition," *IEEE Access*, vol. 10, pp. 97945–97953, 2022.

[54] F. Boutros, P. Siebke, M. Klemt, N. Damer, F. Kirchbuchner, and A. Kuijper, "Pocketnet: Extreme lightweight face recognition network using neural architecture search and multistep knowledge distillation," *IEEE Access*, vol. 10, pp. 46823–46833, 2022.

[55] G. Bae, M. de La Gorce, T. Baltrušaitis, C. Hewitt, D. Chen, J. Valentin, R. Cipolla, and J. Shen, "Digiface-1m: 1 million digital face images for face recognition," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 3526–3535, 2023.

[56] H. Qiu, B. Yu, D. Gong, Z. Li, W. Liu, and D. Tao, "Synface: Face recognition with synthetic data," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10880–10890, 2021.

[57] M. Kim, F. Liu, A. Jain, and X. Liu, "Dcface: Synthetic face generation with dual condition diffusion model," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12715–12725, 2023.

[58] D. Cao, X. Zhu, X. Huang, J. Guo, and Z. Lei, "Domain balancing: Face recognition on long-tailed domains," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5671–5679, 2020.

[59] Z. Zhu, G. Huang, J. Deng, Y. Ye, J. Huang, X. Chen, J. Zhu, T. Yang, D. Du, J. Lu, *et al.*, "Webface260m: A benchmark for million-scale deep face recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[60] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4690–4699, 2019.

[61] B. Maze, J. Adams, J. A. Duncan, N. Kalka, T. Miller, C. Otto, A. K. Jain, W. T. Niggel, J. Anderson, J. Cheney, *et al.*, "Iarpa janus benchmark-c: Face dataset and protocol," in *2018 international conference on biometrics (ICB)*, pp. 158–165, IEEE, 2018.

[62] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao, "Ms-celeb-1m: A dataset and benchmark for large-scale face recognition," in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14*, pp. 87–102, Springer, 2016.

[63] F. Wang, L. Chen, C. Li, S. Huang, Y. Chen, C. Qian, and C. C. Loy, "The devil of face recognition is in the noise," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 765–780, 2018.

[64] A. Nech and I. Kemelmacher-Shlizerman, "Level playing field for million scale face recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7044–7053, 2017.

[65] D. A. Montecino, C. A. Perez, and K. W. Bowyer, "Two-level genetic algorithm for evolving convolutional neural networks for pattern recognition," *IEEE Access*, vol. 9, pp. 126856–126872, 2021.

[66] D. P. Benalcazar, J. E. Zambrano, D. Bastias, C. A. Perez, and K. W. Bowyer, "A 3d iris scanner from a single image using convolutional neural networks," *IEEE Access*, vol. 8, pp. 98584–98599, 2020.

[67] D. R. Vilar and C. A. Perez, "Extracting structured supervision from captions for weakly supervised semantic segmentation," *IEEE Access*, vol. 9, pp. 65702–65720, 2021.

[68] X. He, K. Zhao, and X. Chu, "Automl: A survey of the state-of-the-art," *Knowledge-Based Systems*, vol. 212, p. 106622, 2021.

[69] C. A. Perez, C. A. Salinas, P. A. Estévez, and P. M. Valenzuela, "Genetic design of biologically inspired receptive fields for neural pattern recognition," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 33, no. 2, pp. 258–270, 2003.

[70] B. Zoph and Q. V. Le, "Neural architecture search with reinforcement learning," *arXiv preprint arXiv:1611.01578*, 2016.

[71] E. Real, A. Aggarwal, Y. Huang, and Q. V. Le, "Regularized evolution for image classifier architecture search," in *Proceedings of the aaai conference on artificial intelligence*, vol. 33, pp. 4780–4789, 2019.

[72] T. Hassanzadeh, D. Essam, and R. Sarker, "Evodcnn: An evolutionary deep convolutional neural network for image classification," *Neurocomputing*, vol. 488, pp. 271–283, 2022.

[73] Y. Xie, H. Chen, Y. Ma, and Y. Xu, "Automated design of cnn architecture based on efficient evolutionary search," *Neurocomputing*, vol. 491, pp. 160–171, 2022.

[74] L.-C. Chen, M. Collins, Y. Zhu, G. Papandreou, B. Zoph, F. Schroff, H. Adam, and J. Shlens, "Searching for efficient multi-scale architectures for dense image prediction," *Advances in neural information processing systems*, vol. 31, 2018.

[75] Z. Fan, G. Hu, X. Sun, G. Wang, J. Dong, and C. Su, "Self-attention neural architecture search for semantic image segmentation," *Knowledge-Based Systems*, vol. 239, p. 107968, 2022.

[76] R. Tsukada, L. Zou, and H. Iba, "Evolving deep neural networks for x-ray based detection of dangerous objects," *Deep Neural Evolution: Deep Learning with Evolutionary Computation*, pp. 325–355, 2020.

[77] M. B. Calisto and S. K. Lai-Yuen, "Adaen-net: An ensemble of adaptive 2d–3d fully convolutional networks for medical image segmentation," *Neural Networks*, vol. 126, pp. 76–94, 2020.

[78] G. Li, W. Zhou, W. Chen, F. Sun, Y. Fu, F. Gong, and H. Zhang, "Study on the detection of pulmonary nodules in ct images based on deep learning," *IEEE Access*, vol. 8, pp. 67300–67309, 2020.

[79] T. Tanaka, T. Moriya, T. Shinozaki, S. Watanabe, T. Hori, and K. Duh, "Automated structure discovery and parameter tuning of neural network language model based on evolution strategy," in *2016 ieee spoken language technology workshop (slt)*, pp. 665–671, IEEE, 2016.

[80] V. Passricha and R. K. Aggarwal, "Pso-based optimized cnn for hindi asr," *International Journal of Speech Technology*, vol. 22, pp. 1123–1133, 2019.

[81] Z. Gao, Y. Li, Y. Yang, X. Wang, N. Dong, and H.-D. Chiang, "A gpso-optimized convolutional neural networks for eeg-based emotion recognition," *Neurocomputing*, vol. 380, pp. 225–235, 2020.

[82] A. Rajagopal, G. P. Joshi, A. Ramachandran, R. Subhalakshmi, M. Khari, S. Jha, K. Shankar, and J. You, "A deep learning model based on multi-objective particle swarm optimization for scene classification in unmanned aerial vehicles," *IEEE Access*, vol. 8, pp. 135383–135393, 2020.

[83] Y.-Y. Hong, J. V. Taylar, and A. C. Fajardo, "Locational marginal price forecasting using deep learning network optimized by mapping-based genetic algorithm," *IEEE Access*, vol. 8, pp. 91975–91988, 2020.

[84] S. A. Ali, B. Raza, A. K. Malik, A. R. Shahid, M. Faheem, H. Alquhayz, and Y. J. Kumar, "An optimally configured and improved deep belief network (oci-dbn) approach for heart disease prediction based on ruzzo–tompa and stacked genetic algorithm," *IEEE Access*, vol. 8, pp. 65947–65958, 2020.

[85] W. Jian, Y. Zhou, and H. Liu, "Densely connected convolutional network optimized by genetic algorithm for fingerprint liveness detection," *IEEE Access*, vol. 9, pp. 2229–2243, 2020.

[86] Y. Liu, Y. Sun, B. Xue, M. Zhang, G. G. Yen, and K. C. Tan, "A survey on evolutionary neural architecture search," *IEEE transactions on neural networks and learning systems*, 2021.

[87] Z. Zhong, Z. Yang, B. Deng, J. Yan, W. Wu, J. Shao, and C.-L. Liu, "Blockqnn: Efficient block-wise neural network architecture generation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 7, pp. 2314–2328, 2020.

[88] C. Liu, L.-C. Chen, F. Schroff, H. Adam, W. Hua, A. L. Yuille, and L. Fei-Fei, "Auto-deeplab: Hierarchical neural architecture search for semantic image segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 82–92, 2019.

[89] R. Luo, F. Tian, T. Qin, E. Chen, and T.-Y. Liu, "Neural architecture optimization," *Advances in neural information processing systems*, vol. 31, 2018.

[90] X. Zhang, Z. Huang, N. Wang, S. Xiang, and C. Pan, "You only search once: Single shot neural architecture search via direct sparse optimization," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 9, pp. 2891–2904, 2020.

[91] A. Brock, T. Lim, J. M. Ritchie, and N. Weston, "Smash: one-shot model architecture search through hypernetworks," *arXiv preprint arXiv:1708.05344*, 2017.

[92] H. Jin, Q. Song, and X. Hu, "Auto-keras: An efficient neural architecture search system," in *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pp. 1946–1956, 2019.

[93] H. Larochelle, D. Erhan, A. Courville, J. Bergstra, and Y. Bengio, "An empirical evaluation of deep architectures on problems with many factors of variation," in *Proceedings of the 24th international conference on Machine learning*, pp. 473–480, 2007.

[94] H. Xiao, K. Rasul, and R. Vollgraf, "Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms," *arXiv preprint arXiv:1708.07747*, 2017.

[95] A. Krizhevsky, G. Hinton, *et al.*, "Learning multiple layers of features from tiny images," 2009.

[96] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4690–4699, 2019.

[97] P. Ramachandran, B. Zoph, and Q. V. Le, "Searching for activation functions," *arXiv preprint arXiv:1710.05941*, 2017.

[98] B. Ding, H. Qian, and J. Zhou, "Activation functions and their characteristics in deep neural networks," in *2018 Chinese control and decision conference (CCDC)*, pp. 1836–1841, IEEE, 2018.

[99] V. Albiero, X. Chen, X. Yin, G. Pang, and T. Hassner, "img2pose: Face alignment and detection via 6dof, face pose estimation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 7617–7627, 2021.

[100] X. Zhang, R. Zhao, Y. Qiao, X. Wang, and H. Li, "Adacos: Adaptively scaling cosine logits for effectively learning deep face representations," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10823–10832, 2019.

[101] F. Wang, J. Cheng, W. Liu, and H. Liu, "Additive margin softmax for face verification," *IEEE Signal Processing Letters*, vol. 25, no. 7, pp. 926–930, 2018.

[102] B.-N. Kang, Y. Kim, B. Jun, and D. Kim, "Attentional feature-pair relation networks for accurate face recognition," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5472–5481, 2019.

[103] A. R. F. Da Silva, L. M. Pavelski, L. A. Q. C. Júnior, P. H. D. O. Gomes, L. M. Azevedo, and F. E. F. Junior, "An evolutionary search algorithm for efficient resnet-based architectures: a case study on gender recognition," in *2022 IEEE Congress on Evolutionary Computation (CEC)*, pp. 1–10, IEEE, 2022.

[104] S. C. Hoo, H. Ibrahim, and S. A. Suandi, "Convfacenext: Lightweight networks for face recognition," *Mathematics*, vol. 10, no. 19, p. 3592, 2022.

[105] E. Real, A. Aggarwal, Y. Huang, and Q. V. Le, "Regularized evolution for image classifier architecture search," in *Proceedings of the aaai conference on artificial intelligence*, vol. 33, pp. 4780–4789, 2019.