Utilización de Múltiples Modelos de Redes Neuronales Convolucionales para la Detección de Rostros

Pedro Luis González Roa, Pedro Oscar Pérez Murueta, Benjamín Valdés Aguirre, José Antonio Cantoral Ceballos

***Abstract*— En los últimos años se ha demostrado la complejidad en la realización de algoritmos de detección y reconocimiento de rostros, ya que estos requieren de un alto nivel de abstracción porque existe un alto grado de similitud estructural entre los diferentes rostros. De esta manera, se requiere de la utilización de técnicas de *deep learning*, cómo las *redes neuronales convolucionales* (*CNN* por sus siglas en inglés) para la extración de características o *features* que proporciona una imagen. Estas técnicas requieren de un modelo sumamente complejo y una base de datos de gran tamaño para un entrenamiento exitoso que desempeñe correctamente en diferentes situación de iluminación, calidad de la imagen, perfil de la cara, entre otros. Cumplir con las características previamente mencionadas significa una gran inversión de recursos, la cual sólo empresas de tamaños inmensos son capaces de invertir. Por lo que se propone utilizar una combinación de diferentes *CNN*'s previamente entrenadas con diferentes arquitecturas y bases de datos. Se espera que las diferentes perspectivas o características extraidas por cada modelo se complementen entre sí para tener una representación más completa (aunque más compleja) del rostro a analizar sin la necesidad de entrenar un modelo de complejidad tan grande como la previamente mencionada. Finalmente se utilizará esta representación más compleja para determinar los grupos de imágenes de acuerdo a las diferentes personas en la base de datos.**

I. Antecedentes

I-A. Extracción de Características en Imágenes

El reconocimiento facial se encuentra directamente relacionado con uno de los problemas que ha recibido mucha atención en las últimas tres décadas. El problema de agrupamiento de imágenes, también conocido como *Image Clustering (IC)* por sus siglas en inglés, se centra en obtener una representación numérica sobre una imagen y agruparla con imágenes que tienen representaciones similares. [1]

Métodos de *deep learning*, cómo las *redes neuronales convolucionales*, utilizan una casca de múltiples capas de unidades de procesamiento para extraer características específicas de las imágenes. Estas aprenden diferentes niveles de representación con cada capa de convolución que corresponden a los diferentes niveles de abstracción. Donde las primeras capas son capaces de reconocen y de detectar atributos de bajo nivel, similares a los modelos Gabor y SIFT (diseñados hace décadas); mientras que las capas más externas aprenden un nivel de abstracción más alto. Gracias a la variedad de niveles y filtros utilizados, los modelos de esta índole son capaces de tolerar a cierto nivel variaciones de ángulos, iluminación y calidad de la toma. [2]

I-B. Reconocimiento de Rostros en la Actualidad

La detección de rostros consiste en tres fases principales:

- **Detección de rostro**: Aún cuando la detección de un rostro es trivial para los humanos, en términos de visión artificial no es una tarea sencilla. Esta tarea consiste en dado un vídeo o imágen, detectar y localizar un número desconocido de rostros (incluso si no hay alguno). Esta solución consiste en la segmentación, extracción y verificación de las posibles caras en un ambiente no controlado. [3]
- **Alineación del rostro**: Como segundo paso en este proceso, se alinea el rostro de acuerdo a coordenadas canónicas. [2] Se han presentado varias propuestas, de las cuales las *CNN* han obtenido buenos resultados. Para el proposito de este proyecto, se utilizará la propuesta de Zhang [4]. La cual utiliza una *CNN* para la detección y la alineación de los rostros en la misma secuencia.
- **Reconocimiento**: Este último paso consiste en procesar el rostro obtenido en los pasos anteriores. Primero se procesa el rostro con algoritmos para validar si consiste de un rostro verdadero y no modificado. Esta parte del proceso se conoce como *anti-spoofing*. Después se utiliza técnicas de *deep learning* cómo *CNNs* para extraer las características que distinguen un individuo de otro. Finalmente se compara con cada uno de los registros de la base de datos de las personas identificadas para buscar un emparejamiento. Este proceso puede ser descrito de la siguiente manera: [2]

  $$M(F(P(I_1)), F(P(I_2)))$$

  En donde I_1 e I_2 son las imágenes por procesar, P es la función de preprocesamiento, F es la función para obtener las características del rostro, y M es el cálculo de la distancia entre ambas representaciones y consecuentemente la confirmación si es un emparejamiento.

La extracción de estas representaciones no es un procedimiento trivial. Incluso utilizando las técnicas más avanzadas de *deep learning*, es muy probable que el modelo se vea afectado por cambios en el contexto de la imagen. Estos cambios pueden ser la utilización de artefactos como lentes o cubrebocas; diferentes perfiles de la cara en donde se aprecia porciones importantes del rostro; ó incluso diferentes niveles de iluminación y calidad de la imágen. [5]

II. Trabajos Relacionados

II-A. Combinación de Redes Neuronales Convolucionales Previamente Entrenadas

En previas investigaciones diferentes modelos de *deep clustering algorithm* han demostrado un buen desempeño para pequeñas imágenes. Por otro lado, para obtener resultados exitosos en la agrupación de imágenes de mayor complejidad (cómo imágenes de objetos con estructuras complejas) es necesario la utilización de *CNN*s previamente entrenadas para la extracción de características muy específicas.

Existen una variedad de estos modelos, los cuales tienen un mismo objetivo pero aportan una perspectiva diferente gracias a variaciones dentro de las arquitecturas, funciones de activación y/ó base de datos utilizada para el entrenamiento. Para casos específicos un modelo puede tener mejor desempeño que otro en la obtención de características definitivas, pero en otro contexto puede tener un desempeño inferior. Guérin y coautores en la investigación *'Combining pretrained CNN feature extractors to enhance clustering of complex natural images'* utilizan diferentes modelos *CNN* para obtener resultados más constantes en la tarea de agrupación de imágenes complejas. Esto es gracias a que no se puede saber el modelo óptimo para cualquier situación (al menos de que se prueben todos), y al combinar diferentes modelos se obtiene una representación más completa y compleja. Esto es sacrificando una rápida respuesta a cambio de un resultado más fiable.[1]

II-B. Modelo IE-CNN

An-Pint Song comprobó en su investigación *'Similar Face Recognition Using the IE-CNN Model'* la importancia de tomar múltiples perspectivas sobre una misma imágen al realizar el reconocimiento de rostros. Este artículo menciona un fenómeno dentro de la investigación para el desarrollo de modelos de reconocimiento: siempre se utiliza la parte interna del rostro para el entrenamiento de los modelos. Lo cual no se apega completamente a cómo los humanos procesamos un nuevo rostro. Cuando es difícil determinar la comparación entre individuos, incrementamos nuestro enfoque en características específicas de la cara para disernir entre los dos diferentes rostros. [6] [7] Por lo que Song propone un modelo de *CNN* que utiliza diferentes perspectivas enfocadas en partes estratégicas del rostro para obtener una representación mejorada. La cual obtuvo una mejora en los resultados de alrededor del 5 %. [8]

III. Planteamiento del Problema

Reiterando la problematica con la utilización de *CNN* previamente mencionada la extracción de las características únicas de una cara para el reconocimiento facial es un reto no trivial por múltiples factores. Uno de estos es la posición del rostro, ya que es posible ocultar atributos clave que conllevan a una recolección incompleta de la información necesaria. Así mismo, el uso de artefactos como lentes o cubrebocas llevan a obstrucciones en dicha recolección. Por otro lado, las expresiones faciales pueden alterar la representación numérica, incluso cuando es la misma cara. Finalmente, los últimos factores que afectan son relacionados al ambiente, cómo la iluminación y la calidad de la cámara. [5]

El diseño y entrenamiento de modelos lo suficientemente complejos que puedan desempeñarse correctamente en cualquier tipo de contexto requiere de una gran cantidad de recursos de *hardware* y de una base de datos que incluya varias representaciones por cada uno de estos contextos. Incluso corporaciones gigantes han tenido problemas para resolver esta problematica en todos los casos posibles. Podemos tomar como ejemplo la situación en la que se encontró Apple, cuando su reconocimiento facial en sus dispositivos *iphone* no era capaz de disernir correctamente entre individuos del país chino. [9]

IV. Propuesta

IV-A. Descripción de la propuesta

En esta investigación se propone utilizar múltiples modelos de redes neuronales convolucionales (*CNN*) en combinación para obtener una representación de mayor complejidad pero de mayor precisión. Al igual que el modelo de Guérin [1] se espera un aumento en el tiempo de ejecución del programa, pero se busca resultados constantes durante contextos diferentes provenientes de bases de datos públicas.

Así mismo, para acelerar el proceso de comparación entre imágenes se considera utilizar algoritmos de agrupamiento (*clustering*) para realizar un mapa de las diferentes representaciones para cada rostro. De esta manera no es necesario comparar el rostro desconocido con cada una de las imágenes identificadas; sólo se predicirá el *cluster* a cual pertenece.

IV-B. Concatenación de Diferentes Representaciones (CC)

El acercamiento de concatenación consiste en, cómo su nombre lo dice, concatenar las diferentes representaciones obtenidas de cada modelo. En otras palabras, se agregan dimensiones por cada modelo que proporciona una perspectiva.

Ya que se espera que entre más modelos utilizados la complejidad de tiempo crezca exponencialmente, se utilizará el algoritmo de reducción de dimensiones *Principal Component Analysis (PCA)* para reducir el tiempo de procesamiento.

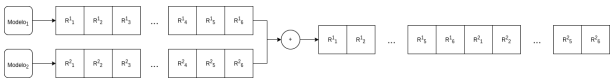

Fig. 1. Representación visual del acercamiento de concatenación.

IV-C. Utilización de Métodos de Consenso de Agrupamiento (MVEC)

Existen diferentes algoritmos e implementaciones para el agrupamiento de información. Aunque haya una gran variedad de estos algoritmos, muchos de estos encuentran soluciones apropiadas pero no óptimas. Estos convergen a un óptimo local y no a un óptimo global, aunque existe la pequeña posibilidad de que algunos sí llegan a este último (cómo por ejemplo: el algoritmo de *k-means*. Ya que llegar a una solución óptima global es un poco aleatoria

por la dependencia de la inicialización y distribución de las diferentes entradas, escoger un algoritmo de agrupación apto para la información persentada puede ser una tarea difícil. Para solucionar este problema, diferentes investigadores introdujeron el concepto de métodos de conjuntos para agrupamientos (*Cluster Ensembles - CE* en inglés), ó también conocidos como métodos de consenso de agrupamiento (*Consensus Clustering* en inglés). Los algoritmos de *CE* combinan los diferentes resultados de agrupamiento para generar un agrupamiento final, sin necesitar acceso a los algoritmos o registros de información. [10]

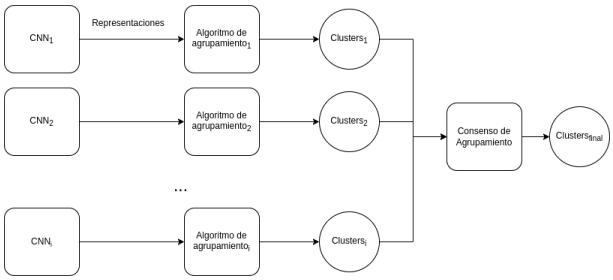

Fig. 2. Representación visual del acercamiento de consenso.

Para el enfoque de este proyecto, utilizaremos la librería de Python proporcionada por Sano Takehiro [11]; con el algoritmo de *Hybrid Bipartite Graph Formulation*. [12]

IV-D. Utilización de Agrupamiento de Múltiples Vistas (MVC)

En la era de *Big Data*, se obtienen perspectivas diferentes de un mismo objeto desde una gran variedad de sensores. Estos sensores producen una salida de diferentes características entre sí, es decir son perspectivas que se complementan entre ellas para una representación más compleja del objeto observado. Por lo cual, ha surgido una tendencia a experimentar con algoritmos que puedan utilizar las diferentes dimensiones de cada vista para hacer predicciones más certeras. En el ámbito de agrupamiento de información de múltiples vistas, ha incrementado en popularidad el algoritmo de Agrupación de vistas múltiples; *Multi-View Clustering (MVC)* en inglés.

Esta exhibición de propiedades heterógenas contiene un potencial de contener posibles conecciones entre ellas, las cuales pueden ser explotadas para desenmascarar características únicas de cada entrada de información. La idea principal de utilizar *MVC* es particionar objetos de acuerdo a diferente criteria relacionada con estas conecciones de sus diferentes vistas. [13]

IV-E. Medición

Para evaluar el desempeño de este modelo de combinación se dividirá en dos etapas dicha evaluación:

1. Se dará la tarea de agrupar las imágenes de un número conocido y variable de personas. La composición de estos grupos será evaluada con la métrica *Normalized Mutual Info (NMI)* [14].
2. De un mapeo previamente realizado de un número conocido de personas, se agregarán más imágenes de personas incluidas en dicho agrupamiento y de otras no incluidas. Por cada imágen se evaluará si predijo correctamente la persona a la que pertenece o si es una persona nueva. Denotando el porcentaje de casos de falsos positivos, ya que es una métrica importante en los sistemas de seguridad biométricos.
3. La diferencia de tiempo para realizar el agrupamiento entre cada uno de los acercamientos, así como también el tiempo que conlleva predecir una nueva imágen.

V. Resultados

VI. Conclusiones

VII. Futuras Aportaciones

References

[1] J. Guérin, S. Thiery, E. Nyiri, O. Gibaru, and B. Boots, "Combining pretrained cnn feature extractors to enhance clustering of complex natural images," *Neurocomputing*, vol. 423, pp. 551–571, 1 2021.

[2] M. Wang and W. Deng, "Deep face recognition: A survey," *Neurocomputing*, vol. 429, pp. 215–244, 3 2021.

[3] E. Hjelmås and B. K. Low, "Face detection: A survey," *Computer Vision and Image Understanding*, vol. 83, pp. 236–274, 2001.

[4] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multi-task cascaded convolutional networks," *IEEE Signal Processing Letters*, vol. 23, pp. 1499–1503, 4 2016.

[5] M. Bodini, "A review of facial landmark extraction in 2d images and videos using deep learning," *Big Data and Cognitive Computing*, vol. 3, pp. 1–14, 3 2019.

[6] A. W. Young, D. Hellawell, and D. C. Hay, "Configurational information in face perception," *Perception*, vol. 16, pp. 747–759, 1987.

[7] T. J. Andrews, J. Davies-Thompson, A. Kingstone, and A. W. Young, "Internal and external features of the face are represented holistically in face-selective regions of visual cortex," *Journal of Neuroscience*, vol. 30, pp. 3544–3552, 3 2010.

[8] A. P. Song, Q. Hu, X. H. Ding, X. Y. Di, and Z. H. Song, "Similar face recognition using the ie-cnn model," *IEEE Access*, vol. 8, pp. 45244–45253, 2020.

[9] G. Birchall, T. Michael, and T. Sun, "Chinese users claim iphone x face recognition can't tell them apart," 12 2017.

[10] K. Golalipour, E. Akbari, S. S. Hamidi, M. Lee, and R. Enayatifar, "From clustering to clustering ensemble selection: A review," *Engineering Applications of Artificial Intelligence*, vol. 104, p. 104388, 9 2021.

[11] T. Sano, "ClusterEnsembles," 8 2021.

[12] X. Z. Fern and C. E. Brodley, "Solving cluster ensemble problems by bipartite graph partitioning," *Proceedings, Twenty-First International Conference on Machine Learning, ICML 2004*, pp. 281–288, 2004.

[13] Y. Yang and H. Wang, "Multi-view clustering: A survey," *Big Data Mining and Analytics*, vol. 1, pp. 83–107, 6 2018.

[14] A. F. McDaid, D. Greene, and N. Hurley, "Normalized mutual information to evaluate overlapping community finding algorithms," 10 2011.