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 extracción 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 sean capaces de complementarse 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 cascada 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 reconocer 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 que han sido evaluadas en previas investigaciones [4], de las cuales las *CNN* han obtenido buenos resultados. Para el propósito de este proyecto, se utilizará la propuesta de Zhang [5]. 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. [6]

II. Trabajos Relacionados

II-A. Combinación de Redes Neuronales Convolucionales Previamente Entrenadas

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, junto con algoritmos de *deep clustering*. [7][8][9]

Existen una variedad de estos modelos *CNN*, 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.[10] 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 discernir entre los dos rostros diferentes.[11][12] 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 %.[10]

III. Planteamiento del Problema

Reiterando la problemática previamente mencionada sobre la utilización de *CNN*, 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 el rostro del mismo individuo. Finalmente, los últimos factores que afectan son relacionados al ambiente, cómo la iluminación y la calidad de la cámara. [6]

El diseño y entrenamiento de modelos lo suficientemente complejos que puedan desempeñarse correctamente en cualquier tipo de contexto requieren 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 problemática 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 discernir correctamente entre individuos del país chino. [13]

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.

Considerando que para determinar si un rostro pertenece a la misma persona, es decir se calcula la distancia entre las representaciones numéricas y se compara en contra de un threshold previamente determinado, se propone 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 y que se mantenga dentro de una distancia definida. En caso de que se encuentre fuera de la distancia definida se considera a la imágen cómo un nuevo individuo a registrar.

Para el desarrollo de esta investigación, se utilizó la librería de Python *Keras-VGGFace*, la cual contiene modelos previamente entrenados con las arquitecturas *VGG16*, *RESNET50*, *SENET50*. Aunque hay más modelos públicos [14][15], se decidió realizar las pruebas con estos tres modelos.

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 evaluará el desempeño del algoritmo de reducción de dimensiones *Principal Component Analysis (PCA)* para reducir el tiempo de procesamiento sin que exista un impacto en la precisión del modelo.

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

Existen diferentes algoritmos e implementaciones para el agrupamiento de información. Aunque haya una gran

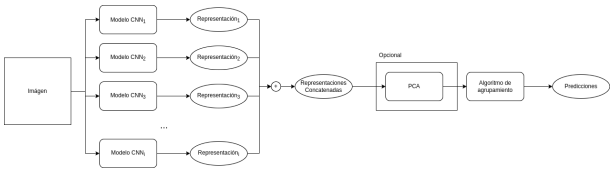

Fig. 1. Representación visual del acercamiento de concatenación.

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 presentada puede ser una tarea difícil. Para solucionar este problema, diferentes investigadores introdujeron el concepto de métodos de conjuntos para agrupamientos (*Cluster Ensembles* - *CE* por sus siglas en inglés), ó también conocidos como métodos de consenso de agrupamiento (*Consensus Clustering*). 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. [16]

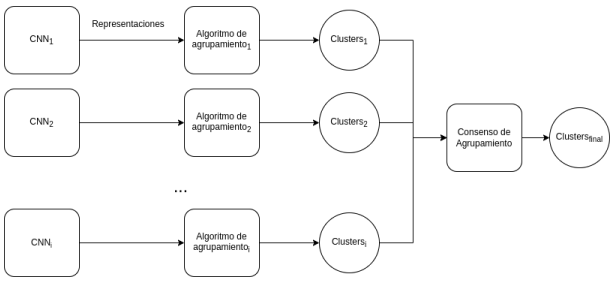

Fig. 2. Representación visual del acercamiento de consenso.

Para el enfoque de este proyecto, utilizaremos la librería de Python proporcionada por Sano Takehiro [17]; con el algoritmo de *Hybrid Bipartite Graph Formulation*. [18]

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. Gracias a la necesidad de identificar diferentes objetos en una gran cantidad de información y vistas, ha incrementado en popularidad los Algoritmos de Agrupación de Vistas múltiples; *Multi-View Clustering (MVC)* en inglés.

Cómo se mencionó previamente, cada vista puede considerarse como un mundo diferente afectado por diferentes variables. Esta exhibición de propiedades heterogéneas contiene un potencial de contener posibles conecciones entre ellas, las cuales pueden ser explotadas para desenmascarar características únicas de cada objeto. La idea principal de utilizar *MVC* es particionar objetos de acuerdo a diferente criteria relacionada con estas conecciones de sus diferentes vistas, ajustando algoritmos comunes y conocidos (cómo *k-means*) para un enfoque de mayor complejidad. [19]

Las pruebas realizadas en esta investigación utilizarán la librería de Python *MVLearn*. [20] La cual contiene mayor variedad de algoritmos de múltiples vistas para utilizar. Cabe mencionar que por las restricciones de este proyecto, sólo se utilizó el algoritmo de *Multi-view K-Means Clustering*.

IV-E. Medición y Datasets

Para evaluar el desempeño de este modelo de combinación, se dará la tarea de 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)* [21]. Es importante esta puntuación porque al evaluar la facilidad del algoritmo para determinar los grupos de acuerdo a las diferentes representaciones significa que estas se encuentran separadas de las representaciones de rostros de otros individuos y también se encuentran cercanas a las representaciones de un mismo individuo; por lo que se cumple nuestra hipótesis de una representación más completa de las características únicas del rostro.

Es importante que estas evaluaciones sean hechas con la mayor cantidad posible de imágenes que representen los diferentes contextos que afectan a este tipo de autenticación biométrica. Por lo que se decidió utilizar cuatro diferentes *datasets* públicas:

- **Yale Face Dataset**: Esta base de datos pública es una de las más comunes y utilizadas para investigación, validación de modelos de detección y reconocimiento de rostros. Esta consiste de un total de 165 imágenes de 15 individuos con diferentes ángulos de iluminación, posición de rostro y utilización de lentes. Este dataset se utilizó para realizar pruebas sencillas durante el desarrollo del modelo propuesto.
- **CelebA Dataset**[22]: Consiste en 10,177 individuos con un total de 202,599 imágenes de rostros. Se recomienda este dataset porque tiene al menos dos o más imágenes por cada una de las personas; lo cual es excelente para probar que las representaciones de un mismo individuo no tengan mucha distancia entre ellas.
- **Labeled Faces in the Wild (LFW)**[23]: Al igual que la base de datos de Yale, *LFW* es muy utilizado en investigaciones. Este *dataset* cuenta con 13,233 imágenes de 5,749 individuos. La desventaja de este dataset es que sólo 1,680 individuos tienen dos o más imágenes.
- **Masked Labeled Faces in the Wild (MLFW)**[24]: Consiste en modificar el *dataset LFW*, agregando una imagen de un cubrebocas para cada rostro. Por lo que

se buscará aprovechar este dataset para probar la consistencia del modelo frente a situaciones inesperadas, cómo lo es una limitación visual para la parte inferior del rostro.

V. Resultados

Al analizar los grandes tamaños de cada *dataset*, se observó que no es viable realizar análisis sobre todas las imágenes de cada uno en un mismo análisis por limitaciones de memoria. En esta investigación buscamos comprobar nuestra hipótesis sobre la utilización de diferentes modelos *CNN* es capaz de construir una representación más completa de un individuo. En otras palabras, buscamos que las representaciones de cada rostros sean sencillas de identificar y agrupar; no buscamos determinar si es posible agrupar una base de datos de imágenes tan extenso. Para solucionar este problema que ocasiona utilizar tanta memoria, optamos por realizar múltiples pruebas de al menos 50 individuos por *dataset* (con la excepción del *Yale dataset* porque cuenta con 15 individuos).

Una vez específicado el tamaño de las pruebas, organizamos la experimentación de este proyecto en dos pasos, cada uno con un objetivo diferente:

V-A. Resultados entre los diferentes algoritmos de agrupación

Cómo primer paso en la implementación de esta investigación se buscó un algoritmo de agrupación que fuera capaz de obtener resultados constantes. Por lo que se realizaron 100 pruebas de alrededor de 50 personas por *dataset* para evaluar tres tipos de algoritmos de agrupación.

- ***k-means***: Se escogió este algoritmo por la idea de que si se encuentra una representación general de un rostro (el *centroide*) es más sencillo encontrar representaciones cercanas a esta representación general de un individuo.
- ***agglomerative***: Como segunda opción se experimentó con *agglomerative clustering* por ser un algoritmo de agrupación jerárquica. Por lo que se tendrá un resultado determinístico.
- ***multiview k-means***: Este es el algoritmo elegido para evaluar el acercamiento de Agrupamiento de Múltiples Vistas.

Ya que se espera la necesidad de trabajar con una gran cantidad de información en muy poco tiempo, también se aprovecharon estas pruebas para determinar si al utilizar un reductor de dimensionalidad se ve afectado el puntaje de los grupos obtenidos.

La figura 3 nos ayuda a visualizar la variación de los resultados preliminares que se obtuvieron para resolver el primer objetivo. Con esta información podemos hacer las siguientes reflexiones:

1. El acercamiento de Agrupamiento de Múltiples Vistas no es viable porque tiende a tener los mismos resultados a cuando se realiza agrupamientos de forma aleatoria. Al analizar los archivos de salida, se puede ver que por la complejidad del algoritmo en todos los casos converge a un mínimo local que no es óptimo porque sólo se logra encontrar una menor cantidad de grupos antes de converger.
2. Existe poca diferencia entre los resultados de los algoritmos *k-means* y *agglomerative*. Por lo que se decidió optar por utilizar el algoritmo de *k-means*, ya que este almacena los *centroides* que se utilizarían en el modelo para determinar si una nueva imagen es parte del grupo de representaciones del mismo rostro.
3. Finalmente utilizar un reductor de dimensionalidad, *PCA*, no altera la precisión al agrupar las imágenes. Por lo tanto, se utilizará este reductor de dimensiones para las siguientes pruebas que contienen un número mucho mayor de imágenes.

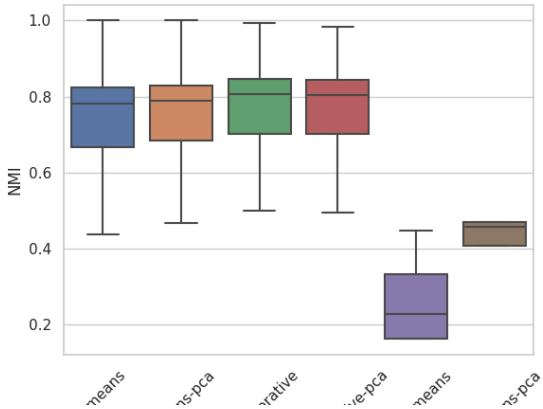

Fig. 3. Análisis de resultados por algoritmo de agrupamiento.

V-B. Resultados de desempeño en agrupación de acuerdo al dataset

Se determinó que nuestro siguiente objetivo era medir la puntuación *NMI* para la agrupación de las imágenes en pruebas de más de 100 individuos. Haciendo énfasis en las diferentes características de cada dataset para realizar reflexiones de estas.

Al reflexionar sobre los diferentes resultados de cada dataset de acuerdo a la figura 5, reflexionamos sobre los siguientes puntos:

1. **Dataset CelebA**: A primera vista los resultados que se obtuvieron en todos los tipos de acercamiento dejaron mucho que desear. Aunque en la figura 6, en la cual se hace una comparación de los resultados cuando se utiliza la arquitectura *VGG16* en combinación con otras, podemos ver que se mantiene el nivel de precisión de este modelo. Al tener un puntaje promedio de 0.82 de forma individual, el modelo *VGG16* proporciona las características más definitivas de los rostros en este dataset. En contraste, las otras dos arquitecturas - *RESNET* y *SENET* - no tienen una perspectiva lo suficientemente clara para determinar con precisión los grupos. Al utilizar el algoritmo de concatenación, las características de la mejor red neuronal son compartidas para que se obtenga un mejor resultado.

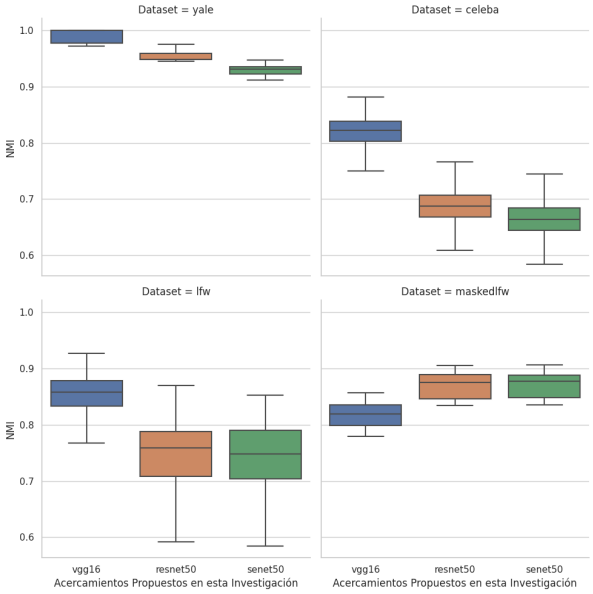

Fig. 4. Análisis de resultados de los modelos sin combinación.

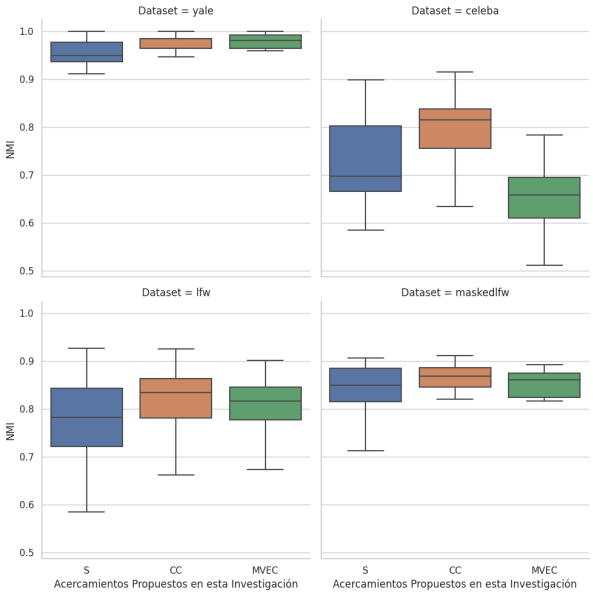

Fig. 5. Análisis de resultados por *dataset*.

2. **Dataset MLFW**: Al analisar la figura 4, nos podemos percatar que es el único dataset en el que la arquitectura *VGG16* no obtuvo los mayores resultados. Este dataset oculta la parte inferior del rostro, por lo que se puede suponer que en esta zona hay características de las que depende esta arquitectura para discernir entre rostros. Aunque en la figura 5, podemos observar que se obtienen mejores resultados gracias a que las otras arquitecturas tienen una perspectiva que complementa a las predicciones de esta.

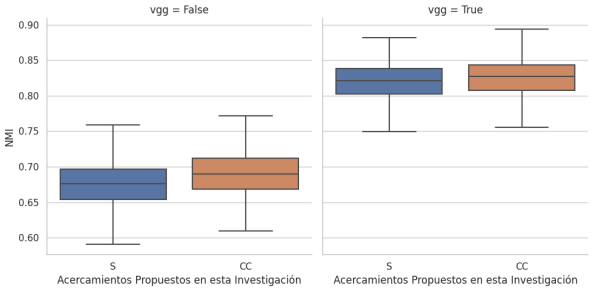

Fig. 6. Análisis de resultados en *celeba* cuando se utiliza VGG.

VI. Conclusiones

En el estado del arte de reconocimiento de rostros se ha popularizado la implementación de *CNN* para la extracción de las características únicas que identifican a un individuo. Esta técnica de *deep learning* se ve afectada directamente por las decisiones arbitrarias del programador al realizar el diseño de esta, por lo que hay una gran variación de resultados (incluso con las mismas arquitecturas) en diferentes casos de prueba. En este artículo se propone la idea de aprovechar las diferentes perspectivas que tienen cada uno de los modelos *CNN* para tener una perspectiva más completa.

Con los resultados y sus reflexiones previamente mencionadas, podemos concluir que en caso de necesitar un modelo que tenga la flexibilidad para la detección de rostros en una gran variedad de contextos es posible utilizar una combinación de redes neuronales previamente entrenadas. Aunque existe la posibilidad de que se seleccione una *CNN* entrenada con un dataset parecido al contexto que se somete y que puede incluso tener el mejor resultado a comparación de otras redes, también existe la posibilidad de elegir una red neuronal que no es apta para la situación que se presenta. Por lo que al sacrificar tiempo para dedicarlo a la ejecución de los diferentes modelos, podemos combinar las diferentes perspectivas para obtener un resultado más constante.

VII. Futuras Aportaciones

En este artículo se realizaron pequeños pero múltiples experimentos para demostrar que es posible combinar diferentes perspectivas en busca de una visión más amplia. Se obtuvieron resultados positivos, incluso cuando sólo se utilizaron tres arquitecturas - *VGG16*, *RESNET50* y *SENET50*. Existe la posibilidad de que se puedan utilizar otras arquitecturas diferentes y encontrar una mayor complementación entre ellas.

Así mismo, cabe mencionar que se experimentó con una pequeña cantidad de algoritmos de agrupamiento; de los cuales existen una mayor variedad (incluso cuando se trata de agrupamiento de múltiples vistas). En un artículo relacionado, Guerín obtuvo resultados muy altos en *datasets* selectos con el algoritmo de *deep clustering* llamado *Joint Unsupervised Learning of Deep Representations and Image Clusters* [8]; con el cual se puede expandir esta experimentación.

Finalmente, se propone experimentar más sobre el comportamiento de estas combinaciones de modelos frente a imágenes nuevas después de haber realizado el agrupamiento inicial. Esto con el objetivo de evaluar a un nivel más profundo las capacidades de autenticación biométrica.

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] S. Zafeiriou, C. Zhang, and Z. Zhang, "A survey on face detection in the wild: Past, present and future," *Computer Vision and Image Understanding*, vol. 138, pp. 1–24, 9 2015.

[5] 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.

[6] 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.

[7] J. Zhao, D. Lu, K. Ma, Y. Zhang, and Y. Zheng, "Deep image clustering with category-style representation," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 12359 LNCS, pp. 54–70, 7 2020.

[8] J. Yang, D. Parikh, and D. Batra, "Joint unsupervised learning of deep representations and image clusters," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2016-December, pp. 5147–5156, 4 2016.

[9] A. F. Agarap and A. P. Azcarraga, "Improving k-means clustering performance with disentangled internal representations," *Proceedings of the International Joint Conference on Neural Networks*, 6 2020.

[10] 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.

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

[12] 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.

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

[14] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2016-December, pp. 2818–2826, 12 2015.

[15] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 07-12-June-2015, pp. 815–823, 3 2015.

[16] 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.

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

[18] 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.

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

[20] R. Perry, G. Mischler, R. Guo, T. Lee, A. Chang, A. Koul, C. Franz, H. Richard, I. Carmichael, P. Ablin, A. Gramfort, and J. T. Vogelstein, "mvlearn: Multiview machine learning in python," *Journal of Machine Learning Research*, vol. 22, no. 109, pp. 1–7, 2021.

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

[22] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

[23] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," Tech. Rep. 07-49, University of Massachusetts, Amherst, October 2007.

[24] C. Wang, H. Fang, Y. Zhong, and W. Deng, "Mlfw: A database for face recognition on masked faces," *arXiv preprint arXiv:2109.05804*, 2021.