Content-Based Image Retrieval

Javier Pérez Vargas, Mario Ruiz Vaquett

***Abstract*—En este proyecto se aborda la Recuperación de Imágenes Basada en el Contenido (CBIR) mediante el análisis de imágenes de paisajes clasificados en cinco categorías: desierto, costa, montaña, glaciar y bosque. Se utilizan cinco métodos distintos de extracción de características: histogramas de color, características de la matriz de co-ocurrencia de niveles de gris (GLCM), características de Histogramas de Gradientes Orientados (HOG), embeddings generados por redes neuronales convolucionales (CNN) y embeddings generados por modelos transformer. Cada método es evaluado utilizando una métrica específica para determinar cuál ofrece el mejor rendimiento en la tarea de recuperación de imágenes.**

I. Introducción

En la era digital, la cantidad de datos visuales ha crecido exponencialmente, presentando desafíos importantes para su organización y recuperación eficiente. La Recuperación de Imágenes Basada en el Contenido (CBIR) se ha consolidado como una técnica innovadora que permite buscar imágenes basándose en sus características visuales, como color, textura, forma y patrones, sin depender de etiquetas o metadatos asociados.

CBIR ha sido un campo de estudio activo durante décadas, evolucionando continuamente para abordar los desafíos planteados por la creciente complejidad de los datos visuales. A lo largo del tiempo, se han desarrollado y aplicado numerosos métodos, como SIFT o SURF, tal y como se describe en [1], mientras que otros utilizan enfoques más modernos, como CNNs o modelos de atención, como se muestra en [2] o [3]. Estos avances han permitido mejoras significativas en la precisión y la eficiencia de los sistemas CBIR, consolidándolos como herramientas clave para la recuperación de información visual en diversos dominios.

El objetivo principal de este proyecto es implementar un sistema CBIR aplicado a paisajes, empleando un conjunto de datos que incluye cinco clases visuales: glaciares, costa, desierto, montaña y bosque. Para ello, se utilizan diferentes métodos de extracción de características, desde técnicas tradicionales como histogramas de color, GLCM y HOG, hasta técnicas avanzadas basadas en aprendizaje profundo, como embeddings de CNN y transformers. El desempeño de cada método será evaluado, lo que permitirá comparar su eficacia en términos de precisión y recuperación. A través de este análisis, se busca identificar el enfoque más efectivo para la tarea de recuperación de imágenes, contribuyendo al desarrollo de sistemas CBIR más eficientes y prácticos.

El código fuente desarrollado para este proyecto está disponible en este repositorio de GitHub.

II. Conjunto de Datos

El conjunto de datos empleado en este trabajo corresponde al *Landscape Recognition Image Dataset*, obtenido de la plataforma Kaggle. Este dataset contiene un total de 12,000 imágenes distribuidas en cinco categorías de paisajes: glaciares, costa, desierto, montaña y bosque. Cada categoría representa un tipo específico de entorno natural, lo que lo hace especialmente útil para tareas de clasificación y recuperación de imágenes basadas en contenido. El conjunto completo puede consultarse en el siguiente enlace: Landscape Recognition Image Dataset.

Para adaptar el conjunto de datos a los objetivos del proyecto, se realizó una selección que permite trabajar con un volumen de datos reducido y balanceado. El conjunto de entrenamiento se compone de 500 imágenes, distribuidas equitativamente con 100 imágenes por cada categoría. Por otro lado, el conjunto de prueba consta de 50 imágenes, con 10 imágenes por categoría. Esta reducción no solo facilita el procesamiento, sino que también asegura un balance adecuado entre las clases, crucial para evaluar con precisión los métodos de recuperación.

III. Métodos Empleados

En este apartado se describen los métodos empleados para la extracción de características visuales de las imágenes, un componente esencial en el diseño de sistemas de recuperación de información basados en contenido. Se han implementado tanto enfoques tradicionales basados en histogramas como métodos avanzados que aprovechan el potencial de las redes neuronales profundas y el mecanismo de atención. Cada método ha sido seleccionado con el objetivo de capturar aspectos complementarios de las imágenes, desde propiedades cromáticas y de textura hasta características geométricas y semánticas de alto nivel.

A. *Histogramas de Color*

Los histogramas de color son un descriptor clásico ampliamente utilizado en la caracterización global de imágenes. Este método mide la distribución de colores en el espacio RGB, generando un vector que describe la composición cromática de la imagen. Para cada canal de color, se calculó un histograma independiente utilizando un número fijo de *bins* (en nuestro caso, 8), y posteriormente se calculó el número de píxeles que caían en cada combinación de bins. Así, dados 3 canales y 8 bins, el número de combinaciones es $8^3 = 512$, de forma que esta será la dimensión del vector que utilizaremos para representar cada imagen.

Esta técnica es robusta frente a transformaciones espaciales, como rotaciones y traslaciones, y resulta particularmente útil en tareas donde la composición de color desempeña un papel determinante, como en la categorización de paisajes o imágenes artísticas. Sin embargo, los histogramas de color pueden ser sensibles a cambios de iluminación y sombras, limitando su efectividad en entornos con variaciones lumínicas significativas.

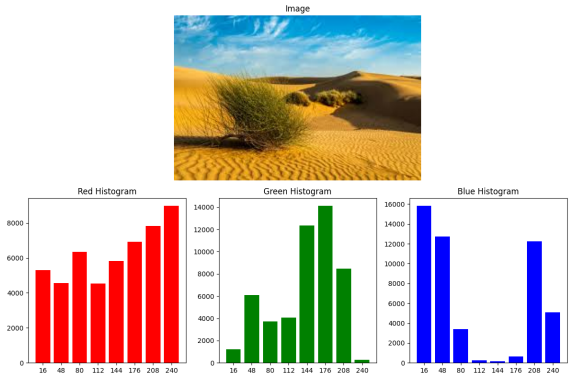

Fig. 1. Histogramas de color, divididos en 8 bins, de cada canal de la imagen.

Además, como es lógico, los histogramas de color, como todas las técnicas basadas en histogramas que veremos, no tienen ninguna noción sobre la semántica de la imagen, que en muchas tareas es esencial. No obstante, para la categorización de paisajes puede aportar un resultado satisfactorio.

B. *Histogramas de Textura basados en GLCM*

La *Gray Level Co-occurrence Matrix* (GLCM) es una herramienta fundamental en el análisis de texturas [4], ya que modela las relaciones espaciales entre pares de píxeles en una imagen. Este enfoque captura patrones estructurales locales al calcular la distribución conjunta de niveles de gris en los píxeles separados por una distancia y orientación específicas, y su uso ha demostrado ser útil en diversos ámbitos [5]. En este trabajo, se generaron las GLCM para todas las combinaciones entre distancias: $\{1, 3, 5, 7\}$ y orientaciones: $\{0°, 45°, 90°, 135°\}$. Esto permitió capturar las dependencias direccionales en las texturas de las imágenes analizadas. Estas orientaciones fueron seleccionadas estratégicamente para garantizar una representación total de las texturas, al abarcar tanto patrones horizontales como diagonales y verticales.

A partir de las matrices obtenidas, se calcularon varias características estadísticas diseñadas para resumir las propiedades más relevantes de las texturas, y las definimos a continuación:

- **Contraste**: Mide la variación local en la intensidad de los píxeles, proporcionando información sobre la rugosidad o suavidad del patrón de textura.
- **Correlación**: Cuantifica el grado de relación lineal entre los valores de intensidad de los píxeles, lo que refleja la periodicidad y la orientación predominante en la textura.
- **Energía**: Una métrica relacionada con la uniformidad, que se emplea para identificar texturas homogéneas o regulares
- **Homogeneidad**: Evalúa la proximidad de los elementos de la matriz a su diagonal principal, lo que está asociado con la similitud de valores en los píxeles adyacentes.
- **ASM (*Angular Second Moment*)**: Representa la uniformidad de la distribución de la GLCM, siendo útil para identificar texturas regulares y bien definidas.
- **Disimilitud**: Mide la diferencia promedio en intensidad entre pares de píxeles, lo que aporta un indicador de la heterogeneidad en la textura.

Cada una de las características extraídas ofrece información sobre diferentes propiedades texturales de las imágenes, lo que permite una representación de la imagen que puede ser útil en tareas de recuperación de imágenes.

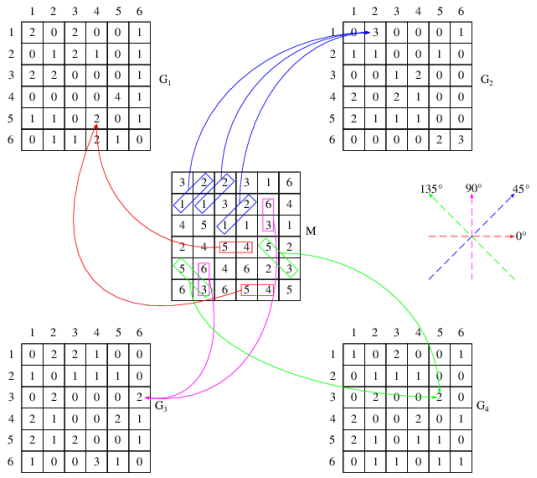

Fig. 2. Obtención de la matriz GLCM para diferentes distancias y orientaciones.

La extracción de estas métricas permite generar descriptores capaces de diferenciar entre texturas complejas, como aquellas encontradas en imágenes de paisajes. Dado que teníamos 4 distancias y 4 orientaciones, obtendremos un total de 16 combinaciones distintas de GLCM. Además, como hemos explicado, cada GLCM aportará 6 métricas distintas. Así, cada imagen vendrá representada por un vector de 96 dimensiones.

Al igual que los histogramas de color, esta técnica no permite capturar la semántica de la imagen, ya que se basa exclusivamente en la representación textural. Esto implica que puede haber limitaciones en su desempeño, especialmente en escenarios donde diferentes clases de paisajes comparten texturas similares, como un desierto y una costa, debido a la presencia de arena. En la sección de resultados se analizará en detalle su rendimiento y se discutirá su efectividad en comparación con otros métodos.

C. *Histograma de Gradientes Orientados*

El *Histogram of Oriented Gradients* (HOG) es un descriptor diseñado para capturar la estructura geométrica y los bordes

de las imágenes. La técnica consiste en dividir la imagen en celdas y calcular un histograma de gradientes orientados para cada una de ellas. Posteriormente, se normalizan los bloques formados por varias celdas para garantizar robustez frente a cambios de iluminación y contraste. Diversos estudios han comprobado su efectividad en campos como la detección de humanos [6], e incluso se ha llegado a combinar con métodos de deep learning [7].

Como se ha mencionado, la extracción de características HOG consiste en dividir la imagen en celdas y describir cada celda. Para que todas las imágenes obtengan una representación de las mismas dimensiones, se han redimensionado a un tamaño de 224×224 píxeles. Además, se han transformado a blanco y negro. Posteriormente, se calcula el descriptor HOG con los siguientes parámetros:

- **Orientaciones = 5**: El número de orientaciones en el histograma de gradientes.
- **Píxeles por celda = (24, 24)**: El tamaño de cada celda, que en este caso es de 24×24 píxeles.
- **Celdas por bloque = (2, 2)**: El número de celdas por bloque, siendo en este caso 2×2.
- **Normalización = 'L2-Hys'**: La normalización de los bloques, que se realiza utilizando la norma L2-Hys, lo que ayuda a mejorar la invariancia frente a cambios de iluminación y contraste.

Estos parámetros aseguran una representación robusta de los bordes y contornos, permitiendo una descripción detallada de las formas en la imagen.

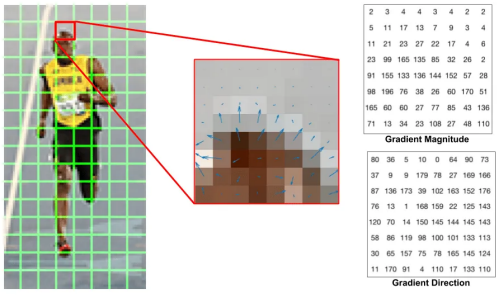

Fig. 3. Obtención de gradientes empleando HOG.

Dada la configuración de parámetros, la imagen de tamaño 224 × 224 píxeles se divide en celdas de 24 × 24 píxeles, resultando en 9 celdas por dimensión (considerando solo celdas completas) y un total de 81 celdas en toda la imagen. Los bloques, formados por 2 × 2 celdas, se deslizan por la imagen con un paso igual al tamaño de una celda, lo que da como resultado 8 bloques por dimensión y un total de 64 bloques. Cada celda genera un histograma con 5 orientaciones, y al incluir 4 celdas por bloque, cada bloque aporta 20 características. Por lo tanto, el vector de características final, compuesto por los 64 bloques, tiene un total de $64 * 20 = 1280$ dimensiones.

Este método es ampliamente reconocido por su capacidad para describir formas y contornos, siendo especialmente útil en la detección de objetos y la representación de imágenes con estructuras geométricas. Al igual que los histogramas de texturas, a priori no podemos determinar cómo será su rendimiento en imágenes de paisajes, pues estos no están tan definidos por formas concretas sino por un conjunto de características más globales.

D. Redes Neuronales Convolucionales (CNN)

Entre los enfoques modernos, hemos considerado utilizar una red neuronal convolucional (CNN). Si bien las CNNs suelen emplearse en tareas como clasificación, detección de objetos o segmentación, también destacan por su capacidad para extraer características visuales y jerárquicas de los datos. Esta propiedad las hace ideales para generar embeddings, es decir, representaciones vectoriales como las que buscamos para comparar imágenes.

Hemos decidido basarnos en arquitecturas bien conocidas y se han valorado opciones como ResNet o DenseNet [8], aunque finalmente hemos utilizado la arquitectura *VGG16* (introducida en 2014 [9]), preentrenada con el conjunto de datos *ImageNet*, pero excluyendo su capa final, permitiendo trabajar directamente con las características generadas por las capas convolucionales. Esta red, conocida por su profundidad y capacidad de extracción de características, puede ser empleada para obtener descriptores muy informativos, usado en múltiples campos, como en medicina [10].

Nuestra modificación a la red consistió en añadir una capa de Global Average Pooling (GAP) para reducir las dimensiones de salida del modelo base, seguida de una capa totalmente conectada (Dense) con 1024 neuronas y función de activación ReLU para capturar representaciones de alto nivel. Se incorporó también una capa Dropout con una tasa del 50% para mitigar el sobreajuste. Finalmente, se añadió otra capa densa de 1024 neuronas con activación softmax para generar los embeddings, obteniendo un modelo completo que transforma las imágenes de entrada en representaciones vectoriales significativas y de alta dimensionalidad (concretamente, 1024 dimensiones).

Para mejorar la capacidad del modelo, decidimos realizar un ajuste fino (*fine-tuning*) en las últimas capas convolucionales de la red base. Inicialmente, las capas base fueron congeladas, es decir, sus pesos se mantuvieron fijos, permitiendo que únicamente las capas añadidas fueran entrenadas. Posteriormente, descongelamos las últimas cuatro capas convolucionales de la red base, permitiendo que sus pesos se reentrenaran y se ajustaran a los patrones específicos de nuestro conjunto de datos. Este ajuste se realizó con una tasa de aprendizaje más baja, asegurando modificaciones graduales en los pesos. Este enfoque de *fine-tuning*, combinado con el diseño de las capas adicionales, permite que el modelo aproveche tanto las características generales aprendidas en *ImageNet* como los patrones específicos de nuestro dominio.

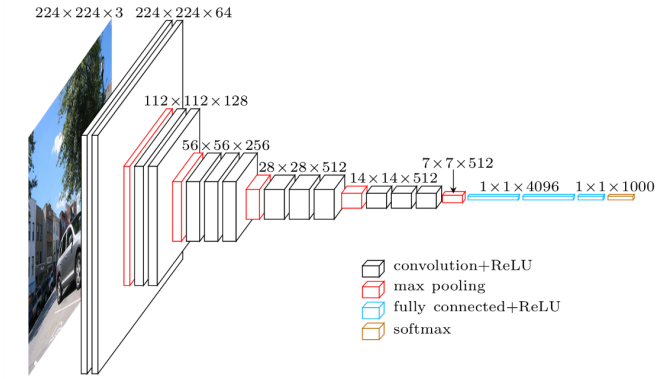

Fig. 4. Arquitectura VGG16 añadiendo capas finales para la generación de embeddings.

Este enfoque permite capturar tanto patrones locales (bordes, texturas) como conceptos más abstractos, ofreciendo una representación que complementa a los descriptores clásicos, y que en este caso sí que puede extraer características semánticas de la imagen, convirtiendo a este método, previsiblemente, en uno de los más efectivos para nuestra tarea.

E. Transformers para Generación de Embeddings

Como método final, hemos aplicado un modelo de *Vision Transformer* (ViT), una arquitectura de última generación basada en transformadores, para la generación de embeddings de alta dimensión. Este modelo segmenta la imagen en parches de tamaño fijo, que son posteriormente proyectados a un espacio vectorial reducido. Estas representaciones son procesadas a través de capas de atención que permiten modelar relaciones espaciales a largo plazo y patrones globales.

Existen numerosos modelos de Vision Transformer (ViT) disponibles, y la gran mayoría son capaces de realizar esta tarea de manera efectiva. En nuestro caso, hemos optado por utilizar el modelo ViT-Base-Patch16-224, desarrollado por Google [11] y preeentrenado con el extenso dataset ImageNet-21k. Este modelo divide las imágenes de entrada, de tamaño 224 × 224 píxeles, en parches de 16 × 16 píxeles que se transforman en vectores mediante una capa de proyección lineal. Este tratamiento, como si fuera una secuencia similar al procesamiento de texto en modelos de lenguaje, permite captar relaciones espaciales a largo plazo. Posteriormente, estos vectores se enriquecen con embeddings posicionales y son procesados por una arquitectura de transformadores basada en múltiples capas de atención. Este enfoque permite al modelo capturar relaciones globales y patrones espaciales complejos, lo que lo convierte en una herramienta poderosa para generar representaciones vectoriales (embeddings) de alta calidad. Concretamente, de 768 dimensiones.

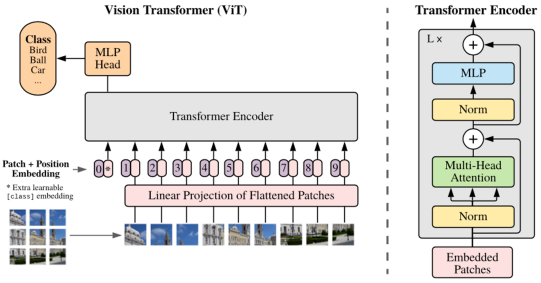

Fig. 5. Arquitectura típica de un ViT, como el utilizado para la generación de embeddings.

El uso de Vision Transformers (ViT) en el sistema CBIR aprovecha su capacidad para procesar información en diferentes escalas y modelar interacciones globales. Esto ofrece mejoras respecto a otras técnicas en tareas como la recuperación de imágenes, al capturar relaciones espaciales más amplias y producir embeddings más representativos.

IV. Indexación

En el sistema de recuperación de imágenes basado en contenido, la indexación de las imágenes se llevó a cabo utilizando *FAISS* (*Facebook AI Similarity Search*), una herramienta diseñada para búsquedas eficientes en espacios vectoriales de alta dimensionalidad. Este proceso garantiza una gestión efectiva de las características extraídas de las imágenes, optimizando la recuperación basada en similitudes.

Las representaciones vectoriales fueron generadas empleando cinco técnicas diferentes, previamente detalladas en la sección anterior: Histogramas de color, Histogramas de textura, *Histogram of Oriented Gradients* (HOG), Redes Neuronales Convolucionales (CNN) y *Transformers* para embeddings. Estas técnicas transforman las propiedades visuales de las imágenes en vectores que capturan información clave como patrones, texturas y características cromáticas.

Tras la extracción de los vectores, se normalizaron empleando la norma L2, un paso crucial para garantizar que todas las representaciones compartan un rango uniforme de magnitudes. Esta normalización mejora la precisión en las métricas de similitud, asegurando que las distancias entre vectores reflejen adecuadamente las diferencias en sus características visuales.

Posteriormente, los vectores fueron indexados utilizando la estructura *FlatL2* de *FAISS*, una implementación eficiente para realizar búsquedas basadas en la distancia euclidiana en espacios vectoriales de alta dimensionalidad. Este índice permite identificar rápidamente las imágenes más similares al consultar el sistema. Además, su almacenamiento persistente facilita la reutilización en futuras operaciones.

V. Evaluación y Resultados

Para evaluar los módelos hemos utilizado la métrica **precision@k**, que mide la precisión de un modelo en las primeras k predicciones. Es especialmente útil en problemas donde los resultados se presentan ordenados según su relevancia, como en sistemas de recomendación o recuperación de información.

Se define de la siguiente manera:

$$\text{precision@k} = \frac{\text{Nº de elementos relevantes entre los } 1^{os}\ k}{k}$$

Donde:

- Los *elementos relevantes* son aquellos que cumplen con el criterio de relevancia para el problema en cuestión. En nuestro caso, serán aquellos que sean de la misma clase (Costa, Desierto...) que la imagen evaluada.
- k es el número de elementos recuperados.

La métrica toma valores entre 0 y 1, donde 1 representa una precisión perfecta en las primeras k predicciones.

Los resultados obtenidos para cada clase y con cada método, utilizando $k = 5$, se pueden observar en la siguiente tabla:

Clase	Color H.	GLCM	HOG	CNN	VIT	Mean
Coast	0.340	0.320	0.140	0.460	**0.940**	0.440
Desert	0.280	0.280	0.280	0.380	**1.000**	0.444
Forest	0.460	0.540	**0.840**	0.640	0.420	0.580
Glacier	0.660	0.380	0.000	0.300	**0.980**	0.464
Mountain	0.240	0.200	0.120	0.300	**0.840**	0.340
Mean	0.396	0.344	0.276	0.416	**0.836**	

TABLE I
RESULTADOS DE PRECISION@K PARA DIFERENTES DESCRIPTORES Y CLASES.

Los resultados presentados en la Tabla I destacan diferencias significativas en el desempeño de los métodos evaluados en un sistema de recuperación de imágenes basado en contenido (CBIR). Entre los cinco enfoques estudiados (Color Histogram, GLCM, HOG, CNN y VIT), VIT (Vision Transformer) se posiciona como el más efectivo, logrando la mejor precisión en casi todas las clases y un promedio general de 0.836. Este resultado resalta la capacidad de VIT para capturar características complejas, lo que lo hace particularmente adecuado para tareas de análisis de imágenes en paisajes variados.

En contraste, HOG (Histogram of Oriented Gradients) muestra el peor desempeño promedio, con un valor de 0.276. Este método se muestra especialmente deficiente en clases como Glacier y Mountain, donde no logra capturar características relevantes, alcanzando una precisión de 0.000 y 0.120, respectivamente. Esto sugiere que los enfoques basados en gradientes y bordes no son efectivos para entender estructuras complejas en imágenes de paisajes naturales. No obstante, y de forma inesperada, HOG es el método que obtiene la precisión más alta en la clase Forest, posiblemente debido a que las imágenes de bosques tienen unos bordes muy representativos.

El desempeño de los métodos tradicionales como Color Histogram y GLCM también es limitado, con promedios de 0.396 y 0.344, respectivamente. Estos métodos ofrecen una representación básica de las imágenes, pero carecen de la capacidad de modelar las relaciones espaciales y contextuales necesarias para discriminar entre clases con características visuales similares. Por otro lado, los métodos basados en aprendizaje profundo, como CNN (0.416) y especialmente VIT, demuestran una ventaja al aprovechar su capacidad para extraer características jerárquicas y contextualizadas de las imágenes.

Los resultados por clase refuerzan la superioridad de VIT, que alcanza su máximo desempeño en categorías como Desert (1.000) y Coast (0.940), donde los paisajes presentan variaciones sutiles y texturas amplias. Esto contrasta con el desempeño mediocre de HOG en estas mismas categorías, evidenciando su incapacidad para modelar detalles complejos.

VI. Conclusiones

Los resultados obtenidos evidencian la capacidad de los enfoques basados en aprendizaje profundo, especialmente Vision Transformers (ViT), en la recuperación de imágenes en entornos complejos y diversos. ViT demostró ser significativamente más efectivo que los métodos tradicionales al capturar tanto características locales como patrones globales, alcanzando la mejor precisión promedio (0.836) y sobresaliendo en clases como Desert y Coast. Esto refuerza la capacidad de las arquitecturas modernas para modelar relaciones espaciales y semánticas más complejas, lo cual es esencial en tareas de recuperación basada en contenido (CBIR).

Por otro lado, los métodos tradicionales como Histogramas de Color, GLCM y HOG mostraron un desempeño inconsistente y deficiencias significativas en clases donde los paisajes comparten características similares. Estos resultados subrayan la necesidad de adoptar enfoques avanzados en sistemas CBIR para abordar la creciente demanda de precisión y adaptabilidad en dominios visuales variados. En este sentido, el desarrollo de herramientas como FAISS para gestionar representaciones vectoriales de alta dimensionalidad se postula como una gran opción en la implementación de soluciones efectivas.

Además, los resultados obtenidos están influenciados por el tamaño reducido del conjunto de datos, limitado a 100 imágenes por categoría. Este sesgo podría mitigarse empleando un dataset más amplio, lo que permitiría capturar mejor la variabilidad de las clases. En particular, se observaron casos de solapamiento entre categorías, como Forest-Mountain y Glacier-Coast, lo que dificulta su diferenciación. Adicionalmente, un análisis manual reveló

que las imágenes de la clase Forest en el conjunto de entrenamiento corresponden predominantemente a bosques otoñales, mientras que en el conjunto de prueba predominan los bosques verdes. Estas discrepancias deben considerarse al interpretar los resultados de la evaluación y al valorar el desempeño de los modelos. Por lo tanto, una posible mejora podría centrarse en expandir de forma equilibrada el conjunto de datos, lo que permitiría mejorar significativamente la generalización del sistema CBIR.

En conclusión, aunque cada método tiene sus fortalezas y debilidades, los transformers han demostrado un rendimiento superior, consolidándose como la solución más efectiva para la recuperación de imágenes. Este resultado refuerza su posición como el estado del arte en visión por computadora y destaca su potencial en el desarrollo de sistemas CBIR.

REFERENCES

[1] Jain, Sahil, Kiranmai Pulaparthi, and Chetan Fulara. "Content based image retrieval." Int. J. Adv. Eng. Glob. Technol 3.10 (2015): 1251-1258.

[2] Wan, Ji, et al. "Deep learning for content-based image retrieval: A comprehensive study." Proceedings of the 22nd ACM international conference on Multimedia. 2014.

[3] Li, Xiaoqing, Jiansheng Yang, and Jinwen Ma. "Recent developments of content-based image retrieval (CBIR)." Neurocomputing 452 (2021): 675-689.

[4] Mohanaiah, P., P. Sathyanarayana, and L. GuruKumar. "Image texture feature extraction using GLCM approach." International journal of scientific and research publications 3.5 (2013): 1-5.

[5] Mall, Pawan Kumar, Pradeep Kumar Singh, and Divakar Yadav. "Glcm based feature extraction and medical x-ray image classification using machine learning techniques." 2019 IEEE conference on information and communication technology. IEEE, 2019.

[6] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection", 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), San Diego, CA, USA, 2005

[7] Binod Bhattarai, Ronast Subedi, Rebati Raman Gaire, Eduard Vazquez, Danail Stoyanov, "Histogram of Oriented Gradients meet deep learning: A novel multi-task deep network for 2D surgical image semantic segmentation", Medical Image Analysis, Volume 85, 2023

[8] Mascarenhas, Sheldon, and Mukul Agarwal. "A comparison between VGG16, VGG19 and ResNet50 architecture frameworks for Image Classification." 2021 International conference on disruptive technologies for multi-disciplinary research and applications (CENTCON). Vol. 1. IEEE, 2021.

[9] Simonyan, Karen, and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition." arXiv preprint arXiv:1409.1556 (2014).

[10] Sharma, Shagun, et al. "A deep learning based convolutional neural network model with VGG16 feature extractor for the detection of Alzheimer Disease using MRI scans." Measurement: Sensors 24 (2022): 100506.

[11] Dosovitskiy, Alexey. "An image is worth 16x16 words: Transformers for image recognition at scale." 2020.