

# CBIR (Content-Based Image Retrieval) en Monumentos: Un Sistema de Recuperación de Imágenes Basado en Contenido

Lucía Rebolledo, Yaiza Sambade

**Resumen**—En este proyecto se desarrolló un sistema de Recuperación de Imágenes Basado en Contenido (CBIR) utilizando características visuales como color, textura, forma y descriptores extraídos mediante redes neuronales convolucionales (CNN) y SIFT. Este sistema permite realizar búsquedas de imágenes similares basándose en su contenido visual, en lugar de depender de etiquetas o descripciones textuales.

Para su implementación se empleó Python, FAISS para la indexación eficiente, y una interfaz interactiva en Streamlit que facilita su uso. Se trabajó con un conjunto de imágenes etiquetadas sobre monumentos, preprocesadas para extraer descriptores visuales que alimentaron los índices de búsqueda.

El rendimiento del sistema fue evaluado con métricas como Precision@K, Recall@K y F1-Score@K, obteniendo resultados favorables en la recuperación de imágenes relevantes.

## I. INTRODUCCIÓN

En la actualidad, la cantidad de imágenes generadas y almacenadas digitalmente ha crecido exponencialmente, generando la necesidad de desarrollar sistemas eficientes para la organización y recuperación de información visual. Tradicionalmente, estos sistemas han dependido de etiquetas o descripciones textuales, las cuales presentan limitaciones cuando son insuficientes, erróneas o inexistentes. En este contexto, los sistemas de Recuperación de Imágenes Basado en Contenido (CBIR, por sus siglas en inglés) han surgido como una alternativa innovadora, permitiendo realizar búsquedas basadas en las características visuales de las imágenes.

Este proyecto se centra en el desarrollo de un sistema CBIR diseñado específicamente para identificar y recuperar imágenes de monumentos históricos. Para lograrlo, se implementaron descriptores visuales que capturan diferentes aspectos de las imágenes, incluyendo histogramas de color, patrones de texturas, formas a través de Momentos de Hu, y características profundas extraídas mediante redes neuronales convolucionales (CNN) y SIFT (Scale-Invariant Feature Transform). Estos descriptores fueron gestionados mediante FAISS, una herramienta eficiente para búsquedas en espacios de alta dimensionalidad, y se integraron en una interfaz interactiva basada en Streamlit para facilitar su uso.

El objetivo principal es evaluar la capacidad del sistema para recuperar imágenes relevantes de monumentos. Este artículo detalla las etapas clave del desarrollo, presenta los resultados obtenidos y discute las fortalezas y limitaciones del sistema en aplicaciones prácticas.

## II. CONJUNTO DE DATOS

El conjunto de datos utilizado en este proyecto fue extraído de la colección "Wonders of the World Image Classification" disponible en Kaggle.

Este dataset contiene imágenes de monumentos y maravillas arquitectónicas icónicas de todo el mundo, organizadas en categorías específicas. Su estructura y diversidad lo convierten en una excelente fuente para desarrollar y evaluar sistemas de Recuperación de Imágenes Basado en Contenido (CBIR).

### A. Description del Dataset

El conjunto de datos incluye muchas imágenes, pero nosotras hemos seleccionado un total de 110 imágenes organizadas en cinco categorías que corresponden a monumentos icónicos reconocidos a nivel mundial. Cada imagen está clasificada en una de las siguientes categorías: el Coliseo Romano, la Torre Eiffel, las Pirámides de Giza, el Burj Khalifa y el Taj Mahal.

Las imágenes están disponibles en formato estándar como JPG y presentan resoluciones variables, lo que permitió realizar un preprocesamiento previo para homogeneizar las dimensiones y optimizar su uso en el sistema.

### B. Organización del Dataset

Se organizó en dos subconjuntos principales para facilitar su uso en el desarrollo y evaluación del sistema CBIR. El primer subconjunto corresponde al conjunto de entrenamiento, que contiene un total de 22 imágenes por categoría, sumando 110 imágenes en total. Estas imágenes fueron utilizadas para construir los índices FAISS y extraer los descriptores visuales necesarios, incluyendo color, textura, forma, características obtenidas con redes neuronales convolucionales (CNN) y descriptores SIFT.

El segundo subconjunto es el de prueba, que incluye 3 imágenes por categoría, totalizando 15 imágenes. Este conjunto fue empleado exclusivamente para evaluar el rendimiento del sistema CBIR mediante la realización de consultas visuales.

## III. METODOLOGÍA

El desarrollo del sistema de Recuperación de Imágenes Basado en Contenido (CBIR) se llevó a cabo siguiendo una metodología estructurada que abarcó desde la preparación de datos hasta la implementación y evaluación del sistema. Este apartado describe los pasos realizados y las técnicas utilizadas:

### A. Preprocesamiento de imágenes

Para asegurar la uniformidad y la calidad de los datos utilizados en el proyecto, se implementaron varios pasos de preprocesamiento. En primer lugar, las imágenes del conjunto de datos fueron redimensionadas a un tamaño estándar de 256x256 píxeles, con el objetivo de homogenizar el formato de entrada y facilitar el procesamiento posterior. Además, los valores de los píxeles fueron normalizados, escalándolos al rango [0, 1], lo que permitió una extracción de características visuales más eficiente y consistente.

### B. Extracción de características visuales

Se emplearon diversos métodos para capturar las características visuales de las imágenes, todos implementados en el archivo 'features\_extractor.py'.

- **Color:** Se emplearon histogramas de color para analizar y representar la distribución cromática de las imágenes. Estos histogramas se generaron en el espacio de color HSV, que es más robusto frente a cambios de iluminación en comparación con el espacio RGB.
- **Textura:** Para describir la textura, se utilizaron Patrones Binarios Locales (LBP), un método que analiza relaciones locales entre los píxeles de una imagen. Esta técnica es particularmente eficaz para identificar patrones texturales repetitivos.
- **Forma:** Las características relacionadas con la forma se capturaron utilizando los Momentos de Hu, que son descriptores geométricos invariantes a rotación, escala y traslación. Estos momentos se calcularon a partir de los contornos de los objetos presentes en las imágenes.
- **Redes Neuronales (CNN):** Se utilizó un modelo de red neuronal convolucional preentrenado (VGG16) para extraer características profundas de las imágenes. Este enfoque permitió identificar patrones complejos y abstractos de alto nivel, como texturas detalladas, estructuras o combinaciones de elementos visuales, que son difíciles de capturar mediante métodos tradicionales.
- **SIFT (Scale-Invariant Feature Transform):** Este método permitió detectar puntos clave y calcular descriptores únicos para cada imagen, proporcionando una representación que es invariante a escala, rotación y cambios de iluminación. Además, los descriptores SIFT fueron promediados para generar una representación compacta, facilitando las comparaciones entre imágenes.

Cada uno de estos métodos fue seleccionado por su capacidad de capturar diferentes aspectos visuales, garantizando así un análisis integral de las imágenes en el sistema CBIR.

### C. Construcción de índices FAISS

Para optimizar la recuperación de imágenes en el sistema CBIR, se utilizaron índices FAISS, que permiten realizar búsquedas eficientes en espacios de alta dimensionalidad.

La construcción de estos índices se inició con la creación de un diccionario maestro diseñado para almacenar las características visuales extraídas de las imágenes de

entrenamiento. Este diccionario contiene claves específicas para cada tipo de descriptor visual, como color, textura, forma, CNN y SIFT, cada una asociada a un sub-diccionario que relaciona los nombres de las imágenes con sus respectivas características.

Posteriormente, se construyeron los índices FAISS para cada tipo de descriptor utilizando la métrica de distancia euclidiana (IndexFlatL2), ideal para calcular similitudes en espacios vectoriales. La dimensionalidad de cada índice se definió en función del tamaño de los vectores de características generados por cada método.

Una vez construidos, los índices se rellenaron con los vectores de características del diccionario maestro. Al completar este proceso, los índices FAISS contenían todas las características de las imágenes de entrenamiento, permitiendo realizar búsquedas rápidas y precisas basadas en el contenido visual de las imágenes.

### D. Implementación de la interfaz

La implementación de la interfaz del sistema CBIR comienza con una configuración inicial que garantiza la integración adecuada de los recursos necesarios para realizar búsquedas y evaluaciones. En este paso, se cargan los índices FAISS correspondientes a los descriptores de color, textura, forma, CNN y SIFT, los cuales fueron previamente generados y almacenados. Estos índices permiten realizar búsquedas rápidas y precisas sobre las características visuales de las imágenes. Además, se carga un archivo CSV que contiene los nombres y las etiquetas de las imágenes del conjunto de entrenamiento, facilitando la comparación y evaluación de los resultados obtenidos durante las consultas.

La interfaz proporciona opciones al usuario para configurar los parámetros de cada consulta. En primer lugar, se ofrece la posibilidad de seleccionar uno de los cinco descriptores visuales disponibles (color, textura, forma, CNN o SIFT) en función de las características que se desean priorizar en la búsqueda. También, se añade una pestaña en la que el usuario selecciona la categoría a la que pertenece la imagen subida, lo que nos facilitará el cálculo de las métricas.

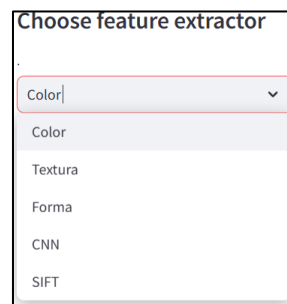


Fig 1. Selección extractor

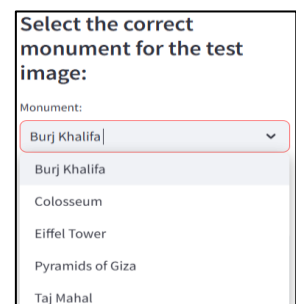


Fig 2. Selección categoría

Asimismo, se habilita una funcionalidad para subir una imagen desde el dispositivo del usuario, que servirá como consulta para buscar imágenes similares.

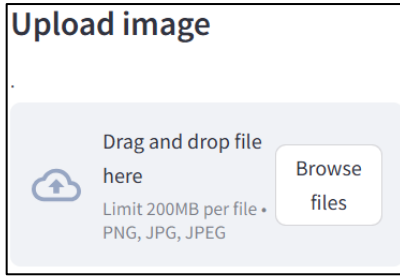


Fig 3. Subir imagen consulta

Antes de realizar la consulta, la imagen puede recortarse utilizando una herramienta interactiva incluida en la interfaz, lo que permite enfocar la búsqueda en áreas específicas de la imagen.

Tras configurar los parámetros de la consulta, el sistema procede a recuperar las imágenes más similares basándose en las características visuales extraídas de la imagen de consulta. Estas características se analizan mediante el descriptor seleccionado y se comparan con los vectores almacenados en los índices FAISS.

#### E. Evaluación del sistema

El rendimiento del sistema se evaluó utilizando métricas estándar:

- Precision@K: Porcentaje de imágenes relevantes entre las K recuperadas.

$$Precisión@K = \frac{True\ Positives}{K}$$

- Recall@K: Porcentaje de imágenes relevantes recuperadas en relación con el total de relevantes disponibles.

$$Recall@K = \frac{True\ Positives}{Total\ relevant}$$

- F1-Score@K: Combinación de precisión y recall para medir el balance entre ambos.

$$F1 - Score@K = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}$$

Estas métricas se calcularon para las diferentes imágenes de test, permitiendo analizar el comportamiento del sistema con cada descriptor.

#### F. Visualización de resultados

Una vez que el usuario sube una imagen de consulta y selecciona el descriptor visual deseado, el sistema realiza una búsqueda en los índices FAISS y recupera las imágenes más similares. Estas imágenes se presentan en un formato de cuadrícula, donde cada resultado incluye una vista previa de la imagen recuperada y su posición en el ranking de similitud. La imagen de consulta se muestra en una posición destacada para facilitar la comparación visual con las imágenes recuperadas.

Para brindar más información sobre el rendimiento del sistema, la interfaz también muestra las métricas de

evaluación correspondientes, como Precision@K, Recall@K y F1-Score@K, calculadas en función de las imágenes recuperadas.

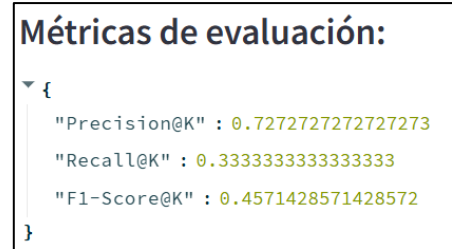


Fig 4. Ejemplo de métricas

Finalmente, el tiempo total de procesamiento de la consulta se muestra al usuario, proporcionando un indicador adicional de la eficiencia del sistema.

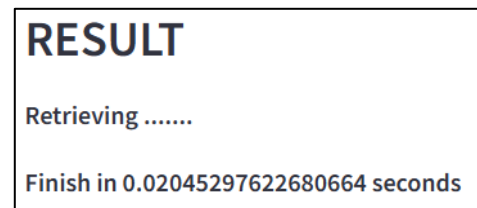


Fig 5. Ejemplo de tiempo de ejecución

## IV. RESULTADOS

### A. Resultados por categoría

- Burj Khalifa: Los mejores resultados para el Burj Khalifa se obtuvieron utilizando los descriptores CNN y SIFT, destacando la importancia de emplear métodos que puedan capturar patrones complejos y puntos clave distintivos en imágenes de estructuras icónicas. Por otro lado, los descriptores basados en color, textura y forma presentaron limitaciones en este caso, debido a las características específicas del edificio y las condiciones de iluminación de las imágenes.
- Torre Eiffel: El descriptor CNN es el más eficaz para identificar imágenes de la Torre Eiffel, debido a su capacidad para capturar patrones complejos y características visuales de alto nivel. Aunque el color, la textura y SIFT también lograron identificar algunas imágenes relevantes, los descriptores basados en forma tuvieron un rendimiento pésimo.
- Coliseo: En este caso, SIFT es el método más eficaz porque su enfoque en puntos clave es ideal para monumentos con detalles estructurales y texturas complejas. Además, los Momentos de Hu ofrecieron un buen rendimiento al capturar las formas geométricas distintivas, CNN también demostró ser efectivo. En contraste, los descriptores basados en textura y los histogramas de color presentaron limitaciones, debido a la diversidad y complejidad visual de las imágenes del monumento.

- Taj Mahal: Los descriptores CNN son los más efectivos para capturar la riqueza visual y los detalles complejos del monumento logrando capturar todas las imágenes relevantes. SIFT también demostró ser un método eficaz, gracias a su capacidad para identificar puntos clave en los detalles arquitectónicos. En contraste, los descriptores basados en color, textura y forma ofrecieron un rendimiento limitado debido a la simplicidad tonal y la uniformidad de las superficies del Taj Mahal, así como a la complejidad arquitectónica que estos métodos no logran capturar completamente.
- Pirámides de Giza: Los descriptores basados en color y CNN demostraron ser los más efectivos para identificar imágenes de las Pirámides de Giza, gracias a la consistencia de las tonalidades y las características visuales complejas que estos métodos son capaces de capturar. En particular, CNN logró recuperar todas las imágenes correspondientes al mismo monumento con gran precisión. Los Momentos de Hu, los descriptores basados en textura y SIFT tuvieron un desempeño más limitado debido a la simplicidad de las texturas y la cantidad moderada de puntos clave distintivos en las imágenes.

#### B. Resultados por extractor

- Color: El histograma de color fue particularmente efectivo para las Pirámides de Giza y, en menor medida, para el Coliseo y la Torre Eiffel. Este método demostró ser más adecuado para monumentos con tonalidades distintivas y consistentes. Sin embargo, su rendimiento se vio afectado por variaciones de iluminación y fondos en imágenes de otros monumentos.
- Textura: El descriptor de textura basado en Patrones Binarios Locales (LBP) mostró limitaciones significativas en la mayoría de los monumentos, con un rendimiento ligeramente mejor en las Pirámides de Giza, donde los patrones texturales de las superficies mostraron algo más de efectividad. Para estructuras con texturas regulares y lisas, como el Burj Khalifa, su desempeño fue bajo, reflejando su incapacidad para capturar detalles complejos.
- Forma: Los Momentos de Hu destacaron principalmente en el Coliseo, donde las formas curvas y los contornos del anfiteatro proporcionaron características geométricas distintivas. En las Pirámides de Giza y el Taj Mahal, los resultados fueron razonables, aunque limitados, debido a sus formas simples y simétricas. Sin embargo, este descriptor presentó un rendimiento deficiente para monumentos con estructuras más complejas, como la Torre Eiffel y el Burj Khalifa.
- CNN: El descriptor basado en redes neuronales convolucionales (CNN) fue consistentemente el más

efectivo en todas las categorías, sobresaliendo especialmente en el Taj Mahal y las Pirámides de Giza. Su capacidad para capturar patrones visuales complejos y características de alto nivel permitió una identificación precisa incluso en monumentos con detalles arquitectónicos y decorativos intrincados. Aunque computacionalmente es más costoso, su desempeño justificó su uso como uno de los mejores métodos.

- SIFT: El método SIFT mostró resultados variados dependiendo del monumento. Fue extremadamente eficaz en el Coliseo, donde la abundancia de puntos clave permitió una recuperación precisa, y tuvo un buen desempeño en el Taj Mahal y el Burj Khalifa. Sin embargo, su efectividad disminuyó en monumentos como las Pirámides de Giza y la Torre Eiffel, donde los puntos clave eran más escasos o menos distintivos. A pesar de estas limitaciones, SIFT sigue siendo una herramienta robusta para identificar imágenes con detalles arquitectónicos y texturales complejos.

#### C. Resultados totales

En general, los resultados del sistema CBIR muestran un desempeño variable dependiendo del monumento y del tipo de descriptor visual utilizado.

Los descriptores basados en redes neuronales convolucionales (CNN) destacaron como los más consistentes y efectivos, logrando identificar imágenes relevantes en todas las categorías gracias a su capacidad para capturar patrones complejos y características de alto nivel.

SIFT también mostró un excelente rendimiento en monumentos con detalles estructurales y texturales distintivos, como el Coliseo Romano y el Burj Khalifa, aunque presentó limitaciones en estructuras más simples como las Pirámides de Giza.

Los descriptores basados en color fueron altamente efectivos en monumentos con tonalidades consistentes y características, como las Pirámides de Giza, pero su rendimiento disminuyó en categorías con variaciones de iluminación y fondos más diversos, como el Burj Khalifa y el Taj Mahal.

Por otro lado, los descriptores de textura y forma ofrecieron un desempeño más limitado, sobresaliendo solo en casos específicos donde las características visuales eran particularmente adecuadas para estos métodos, como las texturas de las Pirámides y las formas del Coliseo.

En conjunto, el sistema demostró ser robusto para recuperar imágenes relevantes, con los descriptores CNN y SIFT como las opciones más confiables y versátiles.

#### D. *Ánàlisis de métricas*

A continuación, se presenta un análisis centrado en la precisión, pues es la métrica que nos aporta la información más relevante. Es decir, nos revela el número de imágenes correctas de todas las totales que te saca la aplicación.

TABLE I. MÉTRICAS

Monument	Color	Textura	Forma	CNN	SIFT
Burj Khalifa	0.27	0.09	0.09	0.72	0.81
Eiffel Tower	0.36	0.27	0	0.81	0.36
Colosseum	0.36	0.27	0.72	0.63	0.90
Taj Mahal	0.27	0.27	0.27	1	0.72
Pyramids of Giza	0.90	0.36	0.36	1	0.18

Analizando estas métricas comprobamos que todo lo anterior se cumple, sobresaliendo el descriptor CNN como el más adecuado para la mayoría de los casos, seguido de SIFT.

#### V. CONCLUSIONES

En este proyecto, se diseñó e implementó un sistema de Recuperación de Imágenes Basado en Contenido (CBIR)

enfocado en la identificación de monumentos, utilizando descriptores visuales de color, textura, forma, redes neuronales convolucionales (CNN) y SIFT. Los resultados obtenidos destacan que los descriptores CNN y SIFT fueron las herramientas más efectivas para capturar patrones visuales complejos y detalles arquitectónicos distintivos, demostrando un rendimiento consistente en casi todas las categorías analizadas. Por otro lado, los descriptores basados en color fueron efectivos en monumentos con tonalidades uniformes y características, mientras que los descriptores de textura y forma ofrecieron un rendimiento más limitado.

En resumen, el sistema CBIR desarrollado, al integrar múltiples descriptores, logró una solución robusta y versátil para la recuperación de imágenes relevantes, permitiendo evaluar las fortalezas y limitaciones de cada enfoque.

#### REFERENCES

- [1] Balabaskar. Wonders of the World - Image Classification [Dataset].Kaggle. <https://www.kaggle.com/datasets/balabaskar/wonders-of-the-world-image-classification>
- [2] Meta AI. FAISS: A library for efficient similarity search and clustering of dense vectors [Software]. Meta AI. <https://ai.meta.com/tools/faiss/>
- [3] Fast Forward Labs. ConvNet Playground [Software]. <https://convnetplayground.fastforwardlabs.com/#/>