EDUCACIÓN
SECRETARÍA DE EDUCACIÓN PÚBLICA

TECNOLÓGICO NACIONAL DE MÉXICO

Tecnológico Nacional de México

Centro Nacional de Investigación y Desarrollo Tecnológico

Tesis de Doctorado

Descripción de Imágenes Naturales en Clases Semánticas para la Recuperación Basada en Contenido

presentada por

M.C. Kevin Salvador Aguilar Domínguez

como requisito para la obtención del grado de

Doctor en Ciencias de la Computación

Director de tesis

Dr. Raúl Pinto Elías

Codirector de tesis

Dr. Juan Gabriel González Serna

Cuernavaca, Morelos, México. Noviembre de 2023.

cenidet
Centro Nacional de Investigación y Desarrollo Tecnológico

SEP TecNM CENTRO NACIONAL DE INVESTIGACIÓN Y DESARROLLO TECNOLÓGICO
RECIBIDO
07 NOV 2023
SUBDIRECCIÓN ACADÉMICA

ESC\FORDOC09

Cuernavaca, Morelos, 26/octubre/2023

ASUNTO: ACEPTACIÓN DEL TRABAJO DE TESIS DOCTORAL

MARÍA YASMÍN HERNÁNDEZ PÉREZ
JEFA DEL DEPARTAMENTO DE CIENCIAS COMPUTACIONALES
PRESENTE

Los abajo firmantes, miembros del Comité Tutorial de la Tesis Doctoral del alumno **KEVIN SALVADOR AGUILAR DOMÍNGUEZ,** manifiestan que después de haber revisado su trabajo de tesis doctoral titulado **"DESCRIPCIÓN DE IMÁGENES NATURALES EN CLASES SEMÁNTICAS PARA LA RECUPERACIÓN BASADA EN CONTENIDO"**, realizado bajo la dirección de Raúl Pinto Elías y la codirección de Juan Gabriel González Serna, el trabajo se ACEPTA para proceder a su impresión.

A T E N T A M E N T E
"Excelencia en Educación Tecnológica®
"Educación Tecnológica al Servicio de México"

RAÚL PINTO ELÍAS
CENIDET

JUAN GABRIEL GONZÁLEZ SERNA
CENIDET

ANDREA MAGADÁN SALAZAR
CENIDET

NIMROD GONZÁLEZ FRANCO
CENIDET

MARÍA YASMÍN HERNÁNDEZ PÉREZ
CENIDET

Osslan Osiris Vergara Villegas

OSSLAN OSIRIS VERGARA VILLEGAS
UNIVERSIDAD AUTÓNOMA DE CD. JUÁREZ

C.c.p.: María Elena Gómez Torres / Jefa del Depto. de Servicios Escolares
Dr. Carlos Manuel Astorga Zaragoza / Subdirector Académico
Expediente

07 NOV 2023

Cuernavaca, Mor., 08/noviembre/2023
No. De Oficio: SAC/172/2023
Asunto: Autorización de impresión de tesis

KEVIN SALVADOR AGUILAR DOMINGUEZ
CANDIDATO AL GRADO DE DOCTOR EN CIENCIAS DE LA COMPUTACIÓN
P R E S E N T E

Por este conducto, tengo el agrado de comunicarle que el Comité Tutorial asignado a su trabajo de tesis titulado **"DESCRIPCIÓN DE IMÁGENES NATURALES EN CLASES SEMÁNTICAS PARA LA RECUPERACIÓN BASADA EN CONTENIDO"**, ha informado a esta Subdirección Académica, que están de acuerdo con el trabajo presentado. Por lo anterior, se le autoriza a que proceda con la impresión definitiva de su trabajo de tesis.

Esperando que el logro del mismo sea acorde con sus aspiraciones profesionales, reciba un cordial saludo.

A T E N T A M E N T E
Excelencia en Educación Tecnológica®
"Conocimiento y tecnología al servicio de México"

EDUCACIÓN | TECNOLÓGICO NACIONAL DE MÉXICO
CENTRO NACIONAL DE INVESTIGACIÓN Y DESARROLLO TECNOLÓGICO
SUBDIRECCIÓN ACADÉMICA

CARLOS MANUEL ASTORGA ZARAGOZA
SUBDIRECTOR ACADÉMICO

C. c. p. Departamento de Ciencias Computacionales
Departamento de Servicios Escolares

CMAZ/lmz

Dedicatoria

Dedico esta tesis a la memoria de mi abuela Cuca Aguilar y a la del doctor Manuel Mejía Lavalle, que no lograron llegar a este momento.

En especial le dedico este logro a esa mujer que me ha enseñado, motivado, y que ha creído en mi incluso cuando yo he dudado. Con la que comparto la pasión por la música, la salud y la ciencia. Aquella que me cuestiona y me hace sentir libere de lo que soy y lo que pienso al mismo tiempo, mi esposa.

Y a aquel amante de la ciencia que me motivo sin saberlo, aquel que paso por el campo, las mesas, la imprenta, para finalmente llegar a las ventas y que cuando parecía su última venta se creó una llave maestra. Ese científico que le toco abrir el camino para que su hijo llegar a la ciencia, gracias, padre

Agradecimientos

Al Consejo Nacional de Humanidades Ciencia y Tecnología (CONAHCYT) por el apoyo económico otorgado para realizar mis estudios de doctorado.

Al Centro Nacional de Investigación y Desarrollo Tecnológico (TecNM/CENIDET), por permitirme utilizar sus instalaciones.

A mi director de tesis el Dr. Raúl Pinto Elías, por su asesoramiento durante el desarrollo de esta tesis, por brindarme sus consejos, su apoyo y su paciencia.

A mis revisores, por su crítica y comentarios que fueron fundamentales para la realización de esta tesis.

Al Dr. Alfredo Cuesta Infante y a todo el laboratorio CAPO de la Universidad Rey Juan Carlos por brindarme espacio, recursos y tiempo durante mi residencia en España.

A mis compañeros y amigos del Departamento de Ciencias Computacionales que me apoyaron con consejos, información, y compañía. En especial al laboratorio de Inteligencia Artificial y con los que he compartido grandes momentos.

A mi esposa y nuestras familias por su consejo, motivación y apoyo durante el desarrollo de este proyecto.

Al caótico universo por brindarme la existencia y la capacidad de la duda.

Resumen

Dentro del campo de la visión por computadora, se han realizado esfuerzos significativos para desarrollar nuevas teorías y modelos de descriptores en sistemas de recuperación de imágenes basados en contenido. Asimismo, se han intentado establecer estandarizaciones para el contenido multimedia, al igual que nuevos descriptores en el estado del arte que incorporan teorías o metodologías que aprovechan la relación entre características de bajo nivel, como el color y la textura. El documento de tesis presenta un análisis y modificaciones de descriptores que utilizan color y textura para la tarea de recuperación de imágenes, además de presentar un nuevo descriptor. El análisis presentado expone algunas de las debilidades presentes en los descriptores clásicos y del estado del arte, que utilizan color y textura, a su vez presenta propuestas e ideas para subsanar las debilidades detectadas. Los experimentos se realizaron utilizando algunos de los conjuntos de imágenes y métricas encontradas en el estado del arte, con la finalidad de realizar un estudio comparativo en relación con el descriptor propuesto. El análisis reveló discrepancias entre el modelo y la implementación de uno de los descriptores, así como los descriptores con mejor rendimiento, sus principales debilidades y las complicaciones al intentar subsanarlos. Durante la evaluación del rendimiento de las variantes propuestas, se detectaron algunas variantes capaces de obtener un rendimiento superior en la recuperación de imágenes en conjuntos con clases semánticas en comparación con descriptores clásicos del estándar MPEG-7 y descriptores recientes del estado del arte. Principalmente la variante llamada *Correlated Microstructure Elements Descriptor* (CMED), que consigue una mejora de hasta 26.41% con respecto al estándar MPEG-7 y 10.75% en comparación al estado del arte. Finalmente, los experimentos y resultados presentados cumplen con los objetivos, alcances y actividades establecidas en el proyecto de investigación.

Palabras clave: Recuperación de imágenes por contenido, Representación de imágenes, Descriptor de Microestructuras, Descriptor de estructuras, Correlación de características visuales, MPEG-7

Abstract

In the field of computer vision, significant efforts have been made to develop new theories and descriptor models in content-based image retrieval systems. Furthermore, attempts have been made to establish standardizations for multimedia content, along with new state-of-the-art descriptors that incorporate theories or methodologies leveraging the relationship between low-level features such as color and texture. The document presents an analysis and modifications of descriptors that use color and texture for image retrieval tasks, along with the introduction of a new descriptor. The analysis presented exposes some weaknesses in both classical and state-of-the-art descriptors that utilize color and texture, while also proposing ideas to address the identified weaknesses. Experiments were conducted using image datasets and metrics found in the state-of-the-art literature, with the aim of conducting a comparative study in relation to the proposed descriptor. The analysis revealed discrepancies between the model and the implementation of one of the descriptors, as well as the descriptors with the best performance, their main weaknesses, and the challenges encountered in attempting to address them. During the evaluation of the performance of the proposed variants, certain variants capable of achieving superior image retrieval performance in datasets with semantic classes were identified, compared to classical descriptors of the MPEG-7 standard and recent state-of-the-art descriptors. Notably, the variant known as the Correlated Microstructure Elements Descriptor (CMED) achieved an improvement of up to 26.41% compared to the MPEG-7 standard and 10.75% compared to the state-of-the-art. Finally, the experiments and results presented in this thesis fulfill the objectives, scope, and activities established in the research project.

Keywords: Content-based image retrieval, Image representation, Microstructure descriptor, Structure descriptor, Visual feature correlation, MPEG-7

Índice

Índice de figuras

Índice de tablas

Lista de acrónimos

ANMRR: *Average Normalized Modified Retrieval Rank*
CBIR: *Content Based Image Retrieval*
CBVIR: *Content-Based Visual Information Retrieval*
CENIDET: *Centro Nacional de Investigación y Desarrollo Tecnológico*
CLD: *Color Layout Descriptor*
CMED: *Correlated Microstructure and Element's Descriptor*
CMED-CF: *Correlated Microestructure Elements Descriptor with Caputo–Fabrizio*
CMED-GL: *Correlated Microestructure Elements Descriptor with Grunwald Letniko*
CMSD: *Correlated Microstructure Descriptor*
CMTH: *Complete Multi-Texton Histogram*
CNN: *Convolutional Neural Network*
ComMSD: *Composite Micro Structure Descriptor*
Corel-CBIR: *Corel Database for Content based Image Retrieval*
DCNN: *Deep Convolutional Neural Network*
DCT: *Discrete Cosine Transform*
EHD: *Edge Histogram Descriptor*
ELM: *Extreme Learning Machines*
GA: *Genetic Algorithm*
GSH: *Gradient Structures Histogram*
HOG: *Histogram of Oriented Gradients*
IEC: *International Electrotechnical Commission*
IR: *Information Retrieval*
ISO: *International Standards Organization*
MAP: *Mean Average Precision*
MDLDPTS: *Multi-Direction and Location Distribution of Pixels in Trend Structure*
MD-TOD: *Multi-Dimensional Texture Orientation Detection*
MIFH: *Multi-Integration Features Histogram*
MIFH-Qm: *Multi-Integration Feature Histogram Quantification maps*
MIFH-Qr: *Multi-Integration Feature Histogram Quantification ranges*
MIFH-R1: *Multi-Integration Feature Histogram Reduced by omissin*
MIFH-R2: *Multi-Integration Feature Histogram Reduced by merge*
MIFH-W: *Multi-Integration Feature Histogram Weighted*
MSD: *MicroStructure Descriptor*
MTH: *Multi Texton Histogram*
MTSD: *Multi-Trend Structure Descriptor*
M-VGG16: *Modified-VGG16*
PCMSD: *Pyramid Correlated Microstructure Descriptor*
QBIC: *Query By Image Content*
ResNet: *Residual Neural Network*
RIPE: *Recuperación de Imágenes por Ejemplo*
SED: *Structure Elements' Descriptor*
SEH: *Structure Elements' Histogram*
TBIR: *Text Based Image Retrieval*
URL: *Uniform Resource Locator*
VGG: *Visual Geometry Group*
XML: *Extensible Markup Language*

1 Introducción

Los sistemas de recuperación de imágenes por contenido "*Content Based Image Retrieval*" (CBIR), se utilizan en diferentes áreas del conocimiento como: Educación[1]–[4], medicina[5]–[10], redes sociales[11]–[13], motores de búsqueda[14], entre otras [15]–[26]. Un ejemplo de CBIR, se encuentra en el buscador web, Google imágenes [27], en su opción de "buscar por imagen", donde se puede seleccionar una imagen desde un archivo o "*Uniform Resource Locator*" (URL), para realizar una búsqueda. Asimismo, a lo largo de los años han surgido ideas de estandarización del contenido multimedia para facilitar el proceso de búsqueda, filtrado y recuperación, como lo es el estándar MPEG-7 propuesto por el grupo "*Moving Picture Experts Group*" (MPEG) [28]. El cual consta de ocho partes orientadas a la descripción del contenido multimedia, es decir a la descripción de audio, imágenes y video.

Sin embargo, aún no se han logrado solucionar del todo las problemáticas presentes en los sistemas de recuperación. Las principales problemáticas en las que se centran las recientes investigaciones van desde la interacción del usuario, segmentación, reducción de dimensionalidad e indexación de características de la imagen, recuperación de imágenes basada en geo etiquetas, características de alto nivel de la imagen, recuperación de imágenes basada en contenido para preservar la privacidad, hasta la recuperación de video basada en contenido [29].

Algunas de las problemáticas en específico las características de alto nivel de la imagen parten de reducir lo que se conoce como, la brecha semántica, que se refiere a la diferencia entre lo que busca un usuario en la recuperación y lo que le entrega el sistema [29], [30]. Para comprender la brecha semántica Eakins establece en [31] tres niveles de recuperación. El primer nivel se refiere a los sistemas que parten de recuperar contenido similar en términos de las características obtenidas directamente de la información de la imagen; como color, textura, forma o la ubicación espacial de los elementos de la imagen. Es decir, la consulta típica podría ser: "busca imágenes como esta" con resultados basados en sus características de bajo nivel. El segundo nivel se centra en sistemas que logran recuperar imágenes similares en relación con la información de objetos dentro de la imagen o estructuras. Por lo que, los sistemas en el nivel dos recuperan objetos de un tipo

determinado identificados por características derivadas, con cierto grado de inferencia lógica. Por ejemplo, “busca una imagen de una flor”. Finalmente, un sistema se considera nivel tres si logra recuperar información relacionada a las características semánticas de la imagen o también conocidas como características de alto nivel. En el nivel 3 se considera que el sistema consigue una recuperación por atributos abstractos, que involucra una cantidad significativa de razonamiento de alto nivel sobre el propósito de los objetos o escenas representadas. Lo que incluye la recuperación de eventos, lugares, actividades, de imágenes con significado emocional, religioso o ideas dentro de la imagen.

En general los usuarios no están únicamente interesados en características de bajo nivel, es decir sistemas nivel uno. Ya que también buscan imágenes relacionadas con características de alto nivel, como lo son actividades, lugares, objetos, emociones, entre otras [32], características que son consideradas en sistemas nivel dos y tres de recuperación. Por lo que, la brecha semántica se establece como la separación que existe entre los resultados obtenidos con los sistemas nivel uno y los sistemas nivel dos y tres.

1.1 Planteamiento del problema

1.1.1 Descripción del problema

Se ha encontrado a lo largo del estado del arte que la brecha semántica puede estar relacionada directamente con las características utilizadas. Ya que la mayoría de los sistemas CBIR utilizan características de bajo nivel que son relativamente fáciles de calcular y, en el mejor de los casos, pueden estar relacionadas con la percepción humana de bajo nivel. Sin embargo, en [33] se menciona que las características de bajo nivel difícilmente estarán directamente relacionados a conceptos de alto nivel. De esta manera, un sistema basado únicamente en características de bajo nivel generará sistemas que logran obtener resultados en el nivel uno de recuperación y difícilmente será capaz de obtener una recuperación en nivel dos o tres.

Por otra parte, las características de alto nivel son difíciles de extraer de los píxeles, considerando que representan conceptos semánticamente significativos en la imagen (por ejemplo, actividades que tienen lugar en la imagen u objetos en la imagen), que son de interés más directo para un ser humano. Extraer características que representen conceptos semánticamente significativos de una imagen, para reducir la brecha semántica, es un desafío para el que no existen soluciones genéricas de alto rendimiento [29].

1.1.2 Delimitación del problema

El presente trabajo se centra en la reducción de la brecha semántica a partir de la obtención de un descriptor que represente características de alto nivel de la imagen. En específico utilizando la relación de características de bajo nivel de la imagen para obtener una mejor evaluación en conjuntos de imágenes que requieran un nivel tres de recuperación.

1.2 Técnicas para la obtención de características de alto nivel

En el estado del arte, con la finalidad de reducir la brecha semántica, se ha intentado extraer características de alto nivel mediante la información encontrada directamente en las

imágenes, es decir, a partir de sus características de bajo nivel. Ya que las características de alto nivel podrían ayudar a la discriminación entre imágenes en términos de semántica. Se han propuesto diversas técnicas para obtener estas características de alto nivel. En [34] identifican los tres siguientes enfoques: (1) utilizar ontologías de objetos para definir conceptos de alto nivel, (2) utilizar métodos de aprendizaje automático para asociar características de bajo nivel con conceptos de consulta, (3) utilizar plantillas semánticas para mapear conceptos de alto nivel con características visuales de bajo nivel. Los tres enfoques son encontrados en el estado del arte; sin embargo, se podría mencionarse un cuarto enfoque. El cual consiste en obtener descriptores que representen semántica de alto nivel derivados de las características de bajo nivel.

En el estado del arte se presentan trabajos [35]–[46], que proponen el uso de descriptores de bajo nivel recientes y clásicos, como los propuestos por el estándar MPEG-7 [47], así como, nuevos descriptores basados en teorías visuales. Los autores utilizan las teorías visuales para mejorar la descripción, relacionando dos o más características de bajo nivel, con la finalidad de obtener descriptores que permitan representar más directamente las características de alto nivel, y así, reducir la brecha semántica [48]. Los descriptores parten de la detección de estructuras y la relación de características del color y textura. Basándose principalmente en la teoría de integración de características presentada por Treisman [49], que establece dos etapas de la visión "*Pre-attentive*" donde se detectan las características de bajo nivel y la etapa "*Attentive*", donde se recombinan las características para formar estructuras más complejas. Asimismo, la teoría de los "*Textons*" de Julesz [50], [51], que presenta el término *Textons* como las microestructuras universales por las que están compuestas las imágenes.

Por otro lado, algunos estudios muestran el uso de redes neuronales profundas en específico el uso de sus características profundas en la tarea de recuperación de imágenes [52]–[55], para poder extraer vectores que representen características semánticas de alto nivel. Sin embargo, es necesario entrenar los modelos con grandes volúmenes de datos, lo que lo limita a las clases o tipos de imágenes utilizadas en el entrenamiento. Además, la longitud del vector de características resulta considerablemente mayor en comparación con enfoques tradicionales.

1.3 Pregunta de investigación

¿Es posible mejorar el rendimiento obtenido mediante los descriptores del estándar MPEG-7 y los del estado del arte para la recuperación de imágenes en nivel tres a partir de un descriptor que considere la relación de sus características de bajo nivel y sus estructuras?

1.4 Descripción de la solución

La solución se buscó en los descriptores basados en teorías visuales, donde se utiliza la relación de dos o más características de bajo nivel y sus estructuras, con la finalidad de obtener descriptores que permitan representar las características de alto nivel y mejorar la recuperación en el nivel tres.

1.4.1 Objetivo General

Investigar, proponer y evaluar un descriptor CBIR mejorado que supere una deficiencia o algunas deficiencias encontradas en los actuales, por medio de un descriptor semántico.

1.4.2 Objetivos Específicos

- Estudiar los descriptores actuales.
- Detectar deficiencias.
- Proponer mejoras.
- Aplicarlo a un dominio aceptado por la comunidad internacional como plataforma de prueba.

1.4.3 Alcances

- El descriptor mejorado deberá superar alguna deficiencia relevante encontrada en descriptores anteriores
- Las pruebas se harán sobre un dominio aceptado internacionalmente y controlado, cuyos resultados correctos sean conocidos de antemano.
- Las pruebas se harán además sobre un dominio del mundo real que tenga impacto en la línea de investigación del grupo de trabajo de Inteligencia Artificial del CENIDET.

1.4.4 Limitaciones

- El tiempo de ejecución no será necesariamente tomado en cuenta.
- Se revisarán descriptores del MPEG-7.
- Se mejorará al menos una de las carencias detectadas.

1.4.5 Justificación

Los sistemas de recuperación actuales requieren de descriptores capaces de representar características de alto nivel, para obtener imágenes similares en términos semánticos. A pesar del esfuerzo realizado por los investigadores, para reducir la brecha semántica utilizando la relación de más de una característica de bajo nivel, o el uso de características profundas, aún falta analizar y mejorar los resultados. El rendimiento de los descriptores en la recuperación se ve afectado por sus debilidades, y las características de bajo nivel utilizadas, lo que evita la reducción de la brecha semántica. En esta dirección, el hecho de proponer descriptores o variantes capaces de subsanar las debilidades presentes en los descriptores actuales ayudaría a obtener un mejor desempeño para la recuperación de imágenes en conjuntos que requieren una recuperación nivel tres. Lo que supondría una reducción de la brecha semántica, acercándonos a sistemas que entreguen resultados esperados por los usuarios.

1.4.6 Beneficios

En este trabajo se propone un descriptor que mejora los resultados de los sistemas CBIR en consultas con clases semánticas. La información obtenida durante la investigación, al igual que los tipos de estructuras y la medida de distancia podrían ser utilizadas en futuras investigaciones. Las áreas beneficiadas van desde la educación, el turismo, los museos, los servicios de investigación, los sistemas de información geográfica, la vigilancia, las compras por internet, y la medicina. Que requieren de un sistema de recuperación de imágenes que considere la semántica. Asimismo, estos avances serían aplicables en otros sistemas de visión artificial que necesiten una representación semántica de la imagen.

1.5 Estructura del documento

El documento continúa con el Capítulo 2, donde se presenta el marco teórico necesario para abordar el tema de investigación, así como los antecedentes y el estado del arte en el campo. En el Capítulo 3, se lleva a cabo un análisis exhaustivo de los descriptores actuales con el objetivo de identificar sus debilidades. El Capítulo 4 describe los procesos seguidos en las propuestas de solución desarrolladas durante el proyecto de investigación. En el Capítulo 5, se detallan los experimentos llevados a cabo para evaluar las propuestas realizadas. El Capítulo 6 muestra las conclusiones obtenidas en la investigación, junto con el trabajo futuro, las contribuciones realizadas y los productos desarrollados a lo largo del proyecto. Por último, se incluyen las referencias bibliográficas y los anexos.

2 Trabajos relacionados

En este capítulo se presentan los conceptos teóricos necesarios para el desarrollo del proyecto de tesis, los trabajos relacionados realizados dentro del Centro Nacional de Investigación y Desarrollo Tecnológico (CENIDET), al igual que los proyectos de investigación presentes en el estado del arte.

2.1 Marco teórico

Con la finalidad de proporcionar la información necesaria para desarrollar el proyecto de tesis, en la siguiente sección se explica brevemente la teoría detrás de los sistemas de recuperación. Así mismo se proporciona información sobre las características de la imagen. Por otro lado, con la finalidad de comprender las técnicas tradicionales y actuales se presenta información sobre el estándar MPEG-7, y las características profundas. Por otra parte, considerando las propuestas realizadas basadas en calculo fraccionario se proporciona una breve introducción al cálculo fraccionario. Finalmente, se detallan las métricas utilizadas en los sistemas de recuperación.

2.1.1 Recuperación de imágenes Basadas en Contenido

La necesidad de recuperar información es tan antigua como el proceso de producción de documentos primitivos. Las primeras colecciones de documentos datan de miles de años. La disponibilidad de documentos escritos finalmente condujo a las primeras colecciones de documentos (archivos y bibliotecas) [56]. Una definición general de lo que es la recuperación de información, "*Information Retrieval*" (IR), proporcionada por [57] es: "encontrar material (generalmente documentos) de naturaleza no estructurada (generalmente texto) que satisfaga una necesidad de información de grandes colecciones (generalmente almacenadas en computadoras)".

La recuperación de información visual está bien fundamentada en el campo de la recuperación de información, que puede describirse como "el proceso de búsqueda (y recuperación) de información relevante dentro de una colección de documentos". La recuperación de información es un campo de investigación maduro y bien establecido. La recuperación de información visual comparte los mismos objetivos que la recuperación de información basada en texto, pero se centra en documentos que contienen información visual, es decir, imágenes y videos [56]. De esta forma surge la recuperación de imágenes

que puede definirse como la tarea de buscar imágenes en un banco de imágenes [58]. La recuperación de imágenes puede dividirse en dos categorías principales, la recuperación de imágenes basadas en contenido CBIR y basadas en texto "*Text Based Image Retrieval*" (TBIR).

Los sistemas de recuperación de imágenes más comunes son los sistemas TBIR, donde la búsqueda se basa en anotaciones automáticas o manuales de imágenes. Las primeras técnicas de búsqueda de imágenes generalmente se basaban en la anotación textual de las imágenes. En otras palabras, las imágenes se anotaron primero con texto y luego se buscaron utilizando un enfoque basado en texto de los sistemas tradicionales de gestión de bases de datos. A través de descripciones de texto, las imágenes se pueden organizar por jerarquías tópicas o semánticas para facilitar la navegación basadas en consultas booleanas estándar. Un sistema TBIR convencional busca en la base de datos el texto similar que rodea la imagen como se indica en la cadena de consulta. En el método de recuperación de imágenes basado en texto, los usuarios usan palabras clave o descripciones de las imágenes como consulta para que puedan usar las imágenes recuperadas, que son relevantes para la palabra clave [29]. La recuperación de imágenes basada en contenido, ampliamente conocida como "*Query By Image Content*" (QBIC) y "*Content-Based Visual Information Retrieval*" (CBVIR), constituye una técnica automatizada en la cual una imagen se utiliza como consulta y, como resultado, se recuperan un conjunto de imágenes similares a la consulta realizada. Este enfoque aprovecha el contenido intrínseco de las imágenes para facilitar el proceso de búsqueda y recuperación, apoyándose en características y patrones visuales en lugar de depender únicamente de metadatos textuales.

Las técnicas de recuperación de imágenes basadas en contenido utilizan contenidos visuales de las imágenes descritas en forma de características de bajo nivel como color, textura, forma y ubicaciones espaciales para representar y buscar las imágenes en las bases de datos. El sistema recupera imágenes similares cuando se presenta una imagen o boceto de ejemplo como entrada al sistema. La imagen de consulta se convierte en la representación interna del vector de características aplicando la misma rutina de extracción de características que se utilizó para construir la base de datos. La medida de similitud se emplea para calcular la distancia entre los vectores de características de la imagen de consulta y los de las imágenes de destino en la base de datos de características. Finalmente, la recuperación se realiza utilizando un esquema de indexación que facilita la búsqueda eficiente de la base de datos de imágenes. Por lo tanto, cuando la medición de similitud se realiza en función de las características de la imagen, el conjunto de salida alcanza un alto nivel de rendimiento de recuperación.

CBIR tiene varias ventajas sobre la recuperación de imágenes tradicional basada en texto. Debido al uso del contenido visual de la imagen de consulta en CBIR, es una forma más eficiente y efectiva de encontrar imágenes relevantes que la búsqueda basada en anotaciones de texto y está más cerca de la percepción humana de los datos visuales. Además, CBIR no consume el tiempo en el proceso de anotación manual como en el

enfoque basado en texto. La representación gráfica de un sistema CBIR se muestra en la Figura 2.1.

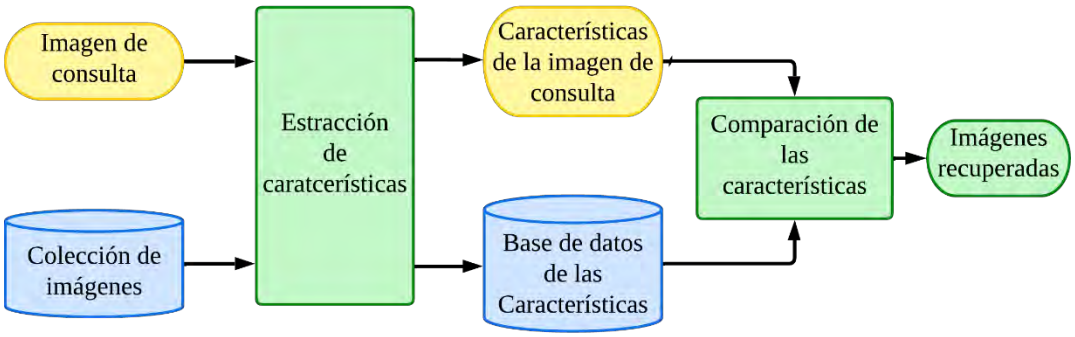


Figura 2.1 Diagrama general de un sistema CBIR [59]

2.1.2 Características de la imagen

Las características de la imagen suelen dividirse en tres niveles: Características de bajo nivel, características de mediano nivel y características de alto nivel. Las características de bajo nivel de una imagen están directamente relacionadas con los aspectos visuales de la imagen. Estas características incluyen color, textura, forma y relación espacial. A continuación, se muestran las tres principales características de bajo nivel de una imagen.

a) Color. El color es el contenido visual más utilizado para la recuperación de imágenes [60], [61]. Sus valores tridimensionales hacen que su potencial de discriminación sea superior a los valores grises unidimensionales de las imágenes. Algunos descriptores de color comúnmente utilizados son los siguientes: los momentos de color, el histograma de color, el vector de coherencia de color y el gráfico de autocorrelación de color [29]. De manera general lo que entrega un descriptor de color es una representación de los valores cromáticos de la imagen.

b) Textura. La textura es otra propiedad importante de las imágenes. Se han investigado varias representaciones de textura en reconocimiento de patrones y visión por computadora. Básicamente, los métodos de representación de texturas se pueden clasificar en dos categorías: estructurales y estadísticos. Los métodos estructurales, describen la textura identificando primitivas estructurales y sus reglas de colocación. Los métodos estadísticos, caracterizan la textura mediante la distribución de la intensidad de la imagen [29]. Los descriptores de textura dan principalmente una representación de información de los bordes es decir los cambios abruptos dentro de la imagen, por ejemplo, donde la imagen tiene cambios de tono, brillo, o saturación.

c) Forma. Las características de forma de objetos o regiones, en comparación con las características de color y textura generalmente se describen después de que las imágenes se han segmentado en regiones u objetos. Dado que la segmentación de imágenes robusta y precisa es difícil de lograr, el uso de características de forma

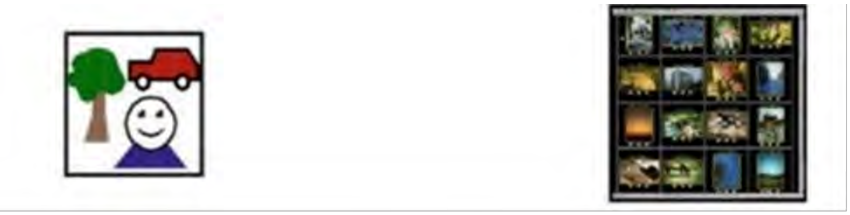
para la recuperación de imágenes se ha limitado a aplicaciones especiales donde los objetos o regiones están fácilmente disponibles. Una buena característica de representación de forma para un objeto debe ser invariable para la traslación, rotación y escala [29].

Las características de bajo nivel representan los aspectos visuales de la imagen, y en el mejor de los casos, pueden estar relacionadas con la percepción humana de bajo nivel. Sin embargo, el contenido visual específico de un dominio, como los rostros humanos, depende de la aplicación y puede involucrar conocimientos previos. Las características de mediano y alto nivel son difíciles de extraer de los píxeles, representan conceptos semánticamente significativos en la imagen (por ejemplo, actividades que tienen lugar en la imagen u objetos en la imagen), que son de interés más directo para un ser humano. Las características de alto nivel se buscan extraer mediante la información encontrada directamente en las imágenes, es decir derivando a partir de sus características de bajo nivel una descripción de un alto nivel semántico. Las características de mediano nivel entonces están relacionadas a objetos dentro de la imagen, lo que bien podría ser, por ejemplo: perro, bolso, auto, flor, por mencionar algunos. Por otra parte, las características de alto nivel se relacionan con los conceptos semánticos que representan la relación o el conjunto de objetos dentro de la imagen, es decir términos muy específicos para un lugar o situación, por ejemplo: África, fiesta, ciudad, bosque, playa, entre otros.

2.1.3 Estándar MPEG-7

MPEG-7, es un estándar desarrollado por la Organización Internacional de Normalización "*International Standards Organization*" (ISO) y la Comisión Electrotécnica Internacional "*International Electrotechnical Commission*" (IEC), especifica una Interfaz de Descripción de Contenido Multimedia. MPEG-7 proporciona una representación estandarizada de metadatos multimedia en "*Extensible Markup Language*" (XML). El objetivo de MPEG-7 es proporcionar un sistema de metadatos interoperables que también esté diseñado para permitir una indexación, búsqueda y filtrado rápido y eficiente de contenido multimedia [47]. Una descripción general del alcance normativo del estándar se muestra en la Figura 2.2 basada en [29].

El estándar MPEG-7 se estructura en siete partes fundamentales: Requisitos del sistema y descripción general, Visual, Audio, Descripción Multimedia Sincrónica, Descripción Multimedia Descriptiva, Descripción de Descripciones Multimedia y Representación de Descripciones Multimedia. La sección visual desempeña un papel clave al proporcionar información sobre los descriptores diseñados para la descripción del contenido visual. Estos descriptores, destinados a extraer características de las imágenes, se dividen en las siguientes categorías Color: *Dominant Color, Scalable Color, Color Layout, Color Structure,* y *Group of Picture Color*; Forma: *Region Shape, Contour Shape*, y *Shape 3D*; Textura: *Homogeneous Texture, Texture Browsing*, y *Edge Histogram*; Movimiento: *Camera Motion, Motion Trajectory, Parametric Motion*, y *Motion Activity*. Además, se incluye un descriptor especializado en la descripción facial denominado *Facial Recognition* [28].



Análisis de contenido	Esquemas de descripción	Búsqueda y filtrado
Extracción de características	Descriptores	Clasificación
Herramientas de anotacíon	Lenguaje	Consulta completa
Autoría	Esquema de codificación	Indexación

Figura 2.2 Descripción general del alcance normativo del estándar MPEG-7

Los descriptores más utilizados en la recuperación de imágenes presentes en el estándar son el descriptor de histograma del borde "*Edge Histogram Descriptor*" (EHD) y el descriptor de diseño del color "*Color Layout Descriptor*" (CLD). El objetivo principal del descriptor CLD es capturar la distribución espacial de los colores en una imagen. Utiliza el espacio de color $YCbCr$ y es muy utilizado por su velocidad e invarianza a la escala y cambios en la resolución. El descriptor se obtiene aplicando la transformada discreta del coseno "*Discrete Cosine Transform*" (DCT) en una matriz 2D de colores representativos locales en el espacio de color $YCbCr$. En la Figura 2.3 se ilustra el proceso de extracción del descriptor basado en [47], que consta de cuatro etapas descritas a continuación:

Paso 1. Se particiona la imagen en 64 (8 × 8) bloques.

Paso 2. Se realiza la detección del color representativo en RGB, utilizando el promedio, dando como resultado un icono de imagen de 64 (8 × 8) colores que son transformados al espacio de color $YCbCr$.

Paso 3. Se aplica la transformada DCT en cada uno de los canales de color del icono de imagen, obteniendo un total de 3 matrices (8 × 8), dando un total de 192 coeficientes DCT.

Paso 4. Finalmente, se realiza una cuantificación no lineal de los coeficientes escaneados en zigzag (64 y 32 niveles para los coeficientes DC y AC respectivamente). El estándar recomienda utilizar un total de 12 coeficientes seis para luminancia (Y), tres para la crominancia (Cb) y tres para (Cr).

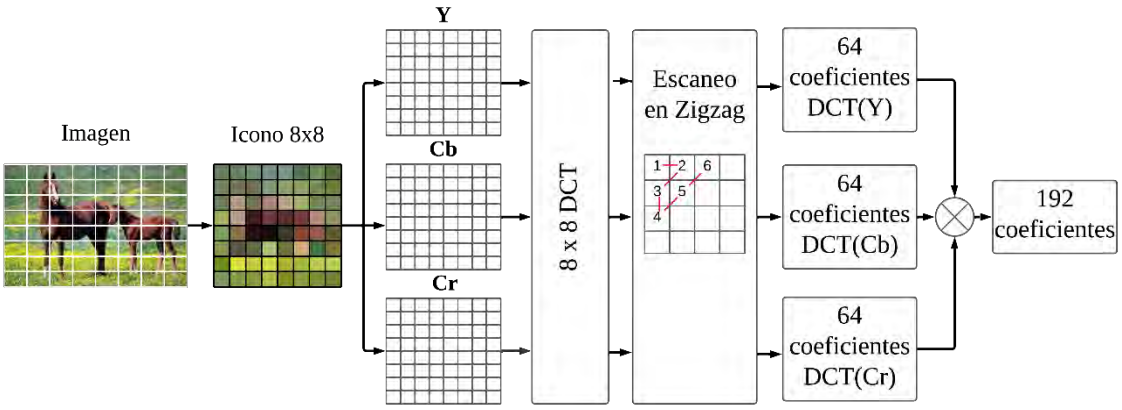


Figura 2.3 Proceso del descriptor CLD

Para medir la similitud entre dos imágenes el descriptor utiliza la medida de distancia mostrada en Eq. (2.1), donde w_{yi}, w_{bi} y w_{ri}, son pesos para ajustar, T la imagen del conjunto de imágenes y Q la imagen de consulta. Siendo entonces, QY_i, QCb_i y QCr_i, los coeficientes i de cada canal Y, Cb y Cr de la imagen de consulta. Asimismo, TY_i, TCb_i y TCr_i, son los coeficientes i de la imagen a comparar.

En el caso del descriptor EHD, su principal objetivo es representar la distribución local de los bordes en una imagen. Trabaja en escala de grises o en el canal Y del espacio de color $YCbCr$, es muy utilizado por sus buenos resultados y velocidad. El proceso que sigue EHD se presenta acontinuacion.

$$L_{CLD}(T,Q) = \sqrt{\sum_i w_{yi} \times (QY_i - TY_i)^2} + \sqrt{\sum_i w_{bi} \times (QCb_i - TCb_i)^2} + \sqrt{\sum_i w_{ri} \times (QCr_i - TCr_i)^2} \quad (2.1)$$

Paso 1. Se particiona la imagen en $16\ (4x4)$ sub-imágenes, y a su vez se dividen en bloques de imagen, como se muestra en la Figura 2.4.

Paso 2. La distribución de borde local para cada sub-imagen se puede representar mediante un histograma. Para generar el histograma, los bordes de los bloques de imagen en las sub-imágenes se clasifican en cinco tipos, vertical, horizontal, 45 ° diagonal, 135 ° diagonal y bordes no direccionales.

Paso 3. Finalmente, se obtiene el histograma para cada una de las 16 sub-imágenes, por lo que se obtiene un total de 5 x 16 = 80 valores en el histograma.

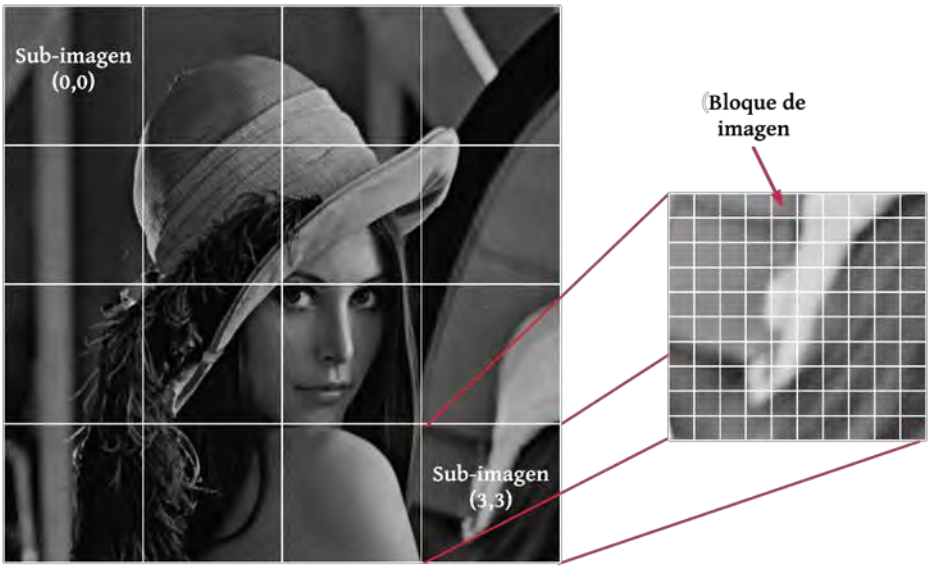


Figura 2.4 Definición de las sub-imágenes y los bloques de imagen en EHD

La medida de distancia utilizada para el descriptor se presenta en la Eq.(2.2), donde $h_Q(i)$ y $h_T(i)$ representa el histograma de cada una de las imágenes, $h_Q^g(i)$ y $h_T^g(i)$ representan el histograma global de la imagen, que acumula los cinco tipos de distribuciones de borde para todas las sub-imágenes, obteniendo 5 valores, de manera similar $h_Q^S(i)$ y $h_T^S(i)$, que representa los histogramas de borde semi globales, agrupando las sub-imágenes en horizontales, verticales y con sus cuatro vecinos. Obteniendo para los histogramas de borde semi globales, un total de (4 + 4 + 5) × 5 = 65 valores.

$$L_{EHD}(T,Q) = \sum_{i=0}^{79} \left| h_Q(i) - h_T(i) \right| + 5 \times \sum_{i=0}^{79} \left| h_Q^g(i) - h_T^g(i) \right| + \sum_{i=0}^{79} \left| h_Q^S(i) - h_T^S(i) \right| \qquad (2.2)$$

2.1.4 Características profundas

Se refiere a las características obtenidas a partir de las capas de convolución de las redes neuronales convolucionales "*Convolutional Neural Network*" (CNN). Las CNN son un tipo de red neuronal profunda que utiliza el proceso de convolución para extraer de manera automática las características visuales de la imagen. El bloque más importante de una CNN es la capa convolucional, a diferencia de las redes neuronales completamente conectadas, las neuronas en la primera capa convolucional no están conectadas a cada píxel en la imagen de entrada, sino solo a los píxeles en sus campos receptivos, es decir su vecindario, dependiendo del tamaño de la ventana de convolución. A su vez, cada neurona en la segunda capa convolucional está conectada solo a las neuronas ubicadas dentro de un pequeño rectángulo en la primera capa. Esta arquitectura permite que la red se concentre en características pequeñas de bajo nivel en la primera capa oculta, luego las ensambla en

características más grandes de nivel superior en la siguiente capa oculta, y así sucesivamente. Esta estructura jerárquica es común en las imágenes del mundo real, que es una de las razones por las que las CNN funcionan para el reconocimiento de imágenes [62].

En la literatura científica, se emplea un enfoque frecuente para capturar características visuales de una imagen utilizando modelos previamente entrenados. En este proceso, se prioriza la utilización de los mapas generados en las capas de convolución del modelo, llegando comúnmente hasta la fase en la que se obtiene un vector plano, conocido como "deep features". Este vector plano se convierte en la entrada para la red neuronal completamente conectada, lo que permite aprovechar las representaciones de alto nivel aprendidas por el modelo previamente entrenado en tareas de clasificación u otras aplicaciones relacionadas con el procesamiento de imágenes.

Entre las arquitecturas más utilizadas en el uso de características profundas se encuentran las redes convolucionales muy profundas propuestas por "*Visual Geometry Group"* (VGG), presentada por Karen Simonyan y Andrew Zisserman de la Universidad de Oxford [63]. La arquitectura de la red comienza con la entrada siendo una imagen de dimensiones (224, 224, 3). Para las capas de convolución, existen cuatro diferentes configuraciones VGG-11, VGG-13, VGG-16 y VGG-19. Cada configuración cambia con respecto a la cantidad de capas de convolución, sin embargo, la arquitectura consta de cinco bloques de convolución, donde las primeras capas tienen 64 filtros de un tamaño de (3, 3), seguido de una capa de (2, 2) con *stride* de dos, el segundo bloque tiene capas de convolución de 128 filtros (3, 3), con *maxpooling* utilizando una ventana de (2, 2) y *stride* de dos. Luego el tercer bloque es compuesto por capas de convolución (3, 3) y 256 filtros. Manteniendo la capa *maxpooling*. Asimismo, el cuarto y quinto bloque son conformados por capas de convolución y una capa de *maxpooling*. Cada capa de convolución (3, 3) con 512 filtros. Finalmente se realiza un aplanamiento de las características profundas para obtener un vector de características. La Figura 2.5 muestra de manera gráfica la arquitectura con la configuración VGG-16 basada en [63].

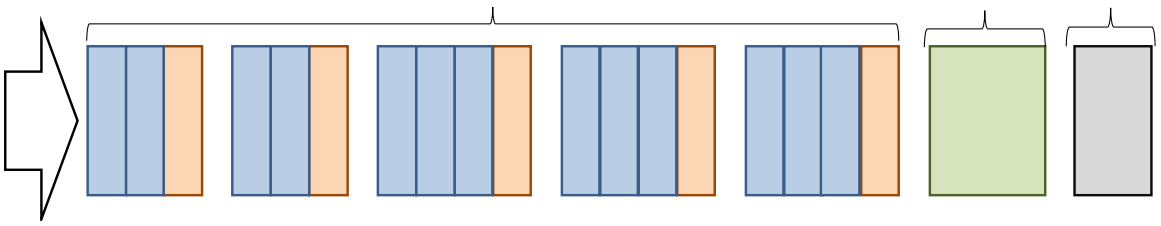


Figura 2.5 Arquitectura VGG-16 [63]

Otra de las arquitecturas más utilizadas en la extracción de características profundas son las redes neuronales residuales "*Residual Neural Network*" (ResNet), propuestas por Shaoqing Ren, Kaiming He, Jian Sun y Xiangyu Zhang en [64]. ResNet a diferencia de las

redes neuronales convolucionales tradicionales utiliza un módulo residual mostrado en la Figura 2.6, con la finalidad de evitar el desvanecimiento del gradiente en el entrenamiento, ya que en redes muy profundas puede llegar a ser muy pequeño. En general existen cinco estructuras de red de profundidad diferentes de 18, 34, 50, 101 y 152, que varían al apilar diferentes números de convoluciones y cantidad de filtros y tamaño. La Figura 2.7 muestra la arquitectura de ResNet-18, basada en [64].

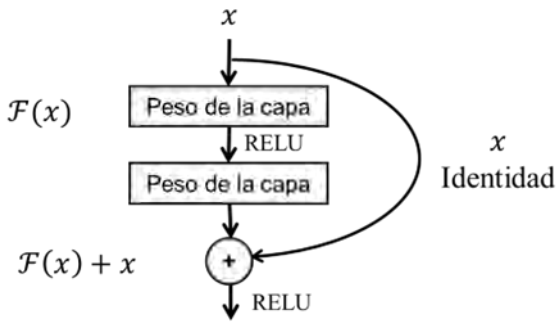


Figura 2.6 Módulo residual [64]

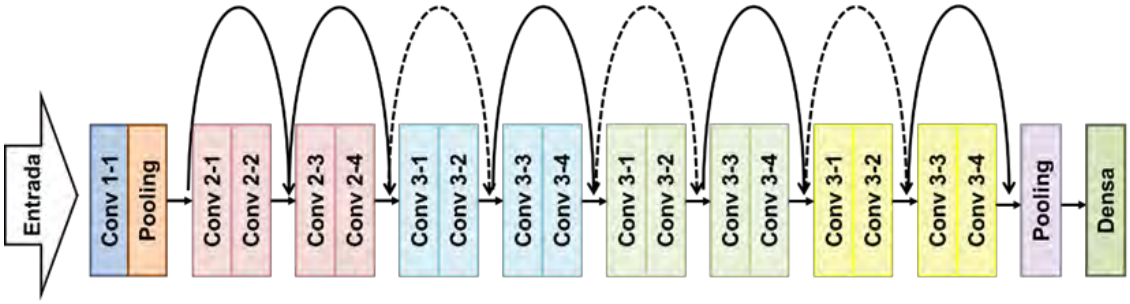


Figura 2.7 Arquitectura de ResNet18 [64]

2.1.5 Cálculo fraccional

Durante la última década métodos no clásico para obtener los bordes de la imagen han surgido, usando técnicas basadas en el cálculo fraccionario o cálculo de orden arbitrario. El cálculo fraccionario tiene sus orígenes en la idea de una derivada de orden no necesariamente entero y aparece en 1695 cuando el marqués de L'Hôpital planteó a Leibniz cuál sería el significado de una derivada de orden $^1/_2$ [65].

En [66] se habla de la historia del cálculo fraccionario. Donde se presenta que, en 1822, Fourier deduce una generalización de los operadores diferenciales e integrales, pero tampoco aportó ninguna aplicación. La primera aplicación conocida del cálculo fraccionario llegó en 1823 por Abel. Donde empleó una derivada de orden ½, y dio una solución tan sencilla y elegante que atrajo la atención de Liouville que en 1832 hizo el primer gran intento de definir la derivada fraccionaria.

En 1847, Riemann escribió un artículo modificando el operador fraccionario dado por Liouville, dando lugar a lo que hoy conocemos como integral fraccionaria de Riemann-Liouville. En la segunda mitad del siglo XIX se destaca a Grünwald que, en 1867, propuso una definición natural y novedosa de derivada e integral de orden arbitrario, y en 1868, Letnikov investigó la derivada de Grünwald y publicó los primeros resultados sobre tal operador. A lo largo del siglo XX con el desarrollo del análisis matemático y la teoría de funciones, aparecen nuevas definiciones de operadores fraccionarios, así en 1917, Hermann Weyl definió una integral fraccionaria adecuada para funciones periódicas, y ya en 1967, Caputo dio una nueva definición de derivada fraccionaria que permitía interpretar físicamente las condiciones iniciales de los problemas. Finalmente, en 1974 se publica el primer texto dedicado enteramente a esta disciplina, "*The Fractional Calculus*", escrito por el físico y matemático J.Spanier y el químico Keith B. Oldham [66].

Entre las definiciones más utilizadas se encuentra la integral fraccionaria de Riemann-Liouville. Fue obtenida por Lauren en 1884 a partir de un contorno dado como un circuito abierto [67]. La derivada de Riemann-Liouville está expresada en la siguiente definición: Sea $\gamma \in \mathbb{R}_+$. El operador ${}^{RL}_{a}\mathcal{D}^{\gamma}_{t}$ está definido sobre $L_1[a, t]$, para $a \leq x \leq t$. Donde $\Gamma(\cdot)$ es la función Gamma de Euler, $n = [\gamma\,]$, finalmente t y τ las variables. El operador ${}^{RL}_{a}\mathcal{D}^{\gamma}_{t}$ está definido por Eq. (2.3).

$$ {}^{RL}_{a}\mathcal{D}^{\gamma}_{t} f(t) = \frac{1}{\Gamma(n-\gamma)} \frac{d^n}{dt^n} \int_a^t \frac{f(\tau)}{(t-\tau)^{\gamma-n+1}} d\tau \tag{2.3} $$

Asimismo, los trabajos de Grünwald [68] y Letnikov [69] sentaron las bases para una nueva definición de derivada fraccionaria. Partieron de la idea de Liouville utilizando el límite de un cociente considerando diferencias de orden fraccionario. El resultado fue la siguiente expresión. Sea $\gamma \in \mathbb{R}_+$. El operador ${}^{GL}\mathcal{D}^{\gamma}_{x}$ está definido por Eq. (2.4), donde $\binom{\gamma}{k}$ es el coeficiente binomial generalizado, $\sum_{k=0}^{\infty}(-1)^k\binom{\gamma}{k} f(x-kh)$ la formulación fraccionaria de una diferencia hacia atrás, h el incremento, y α el coeficiente de derivación.

$$ {}^{GL}\mathcal{D}^{\gamma}_{x} f(x) = \lim_{h \to 0} \frac{\sum_{k=0}^{\infty}(-1)^k\binom{\gamma}{k} f(x-kh)}{h^{\alpha}} \tag{2.4} $$

2.1.6 Métricas de evaluación

La evaluación de la recuperación es de gran importancia y siguiendo la idea de [61] "el objetivo de un sistema de recuperación de imágenes es recuperarlas en un orden de similitud visual", entonces, la evaluación de un sistema CBIR tiene que seguir este criterio. Lo anterior sólo se podría medir si las imágenes a comparar (*ground-truth*) son conocidas. Durante la revisión del estado del arte se encontró únicamente la métrica *Average Normalized Modified Retrieval Rank* (ANMRR) [56], propuesta por el estándar MPEG-7 para la evaluación de sistemas de recuperación. Sin embargo, por encima del 90% de los trabajos revisados usan *Precision* y el resto medidas comúnmente utilizadas para los

sistemas de clasificación. Dentro de los trabajos revisados únicamente se encontraron combinaciones o variaciones de las métricas *Precision* y *Recall*. Por lo que para este trabajo se consideraron las tres métricas más utilizadas en la literatura: *Precision*, *Recall* y *Mean Average Precision* (MAP), así mismo la única métrica encontrada que ha sido propuesta para la evaluación de los sistemas de recuperación, la ANMRR. Las métricas utilizadas son detalladas a continuación:

a) *Precision*. En esta métrica se mide el porcentaje de imágenes correctamente recuperadas en la consulta. La métrica *Precision* puede ser definida como se muestra en la Eq. (2.5). Donde $r(x_n)$ puede tomar valor de cero o uno siguiendo Eq. (2.6). K denota el número de imágenes recuperadas, x_n representa la imagen recuperada en la posición n. Ic_q es el conjunto de imágenes de la categoría correspondiente a la clase de consulta q. Lo que indica que $r(x_n)$ será uno cuando la imagen recuperada en la posición n pertenezca a la misma categoría de la imagen de consulta.

$$P = |K|^{-1} \sum_{n=1}^{K} r(x_n) \tag{2.5}$$

$$r(x_n) = \begin{cases} 1 & si\ x_n \in Ic_q \\ 0 & en\ otro\ caso \end{cases} \tag{2.6}$$

b) *Recall*. Esta métrica establece el porcentaje de imágenes recuperadas pertenecientes a la categoría de la consulta. *Recall* se define en Eq.(2.7), donde R_q es el total de imágenes dentro de la categoría de la imagen de consulta q. Por lo que, *Recall* entrega el porcentaje de imágenes de la categoría que se han logrado recuperar hasta la posición K.

$$R = \left|R_q\right|^{-1} \sum_{n=1}^{K} r(x_n) \tag{2.7}$$

c) *MAP*. La métrica *MAP* puede considerarse una combinación entre *Precision* y *Recall*, obteniendo *Precision* para cada k Eq. (2.8), con relación al total de imágenes en la categoría de la consulta. MAP se obtiene mediante Eq. (2.9)-(2.10), y se podría considerar como el promedio del porcentaje de imágenes correctamente recuperadas para cada posición, con relación a la cantidad de imágenes por recuperar.

$$P_k = |k|^{-1} \sum_{n=1}^{k} r(x_n) \tag{2.8}$$

$$AP_q = \left|R_q\right|^{-1} \sum_{n=1}^{K} P_k \times r(x_n) \quad (2.9)$$

$$MAP = |Q|^{-1} \sum_{q=1}^{Q} AP_q \quad (2.10)$$

d) ***ANMRR***. Presentada por el estándar MPEG-7, *ANMRR* es propuesta para evaluar los sistemas de recuperación considerando la posición en la que cada imagen relevante es recuperada. *ANMRR* entrega un porcentaje de error, por lo que a diferencia del resto de métricas una mejor evaluación es aquel valor más cercano a cero. *ANMRR* inicia asignando un $Rank_n$ a cada imagen x_n perteneciente a la categoría de la consulta Eq. (2.11). Cada imagen relevante se asignada de acuerdo con la posición k en la que es recuperada y se castiga a las imágenes que han sido recuperadas en una posición lejana. El castigo se genera mediante un coeficiente C_q, que considera la cantidad de imágenes que pose cada categoría, como se muestra en Eq. (2.12). Una vez obtenido el $Rank_n$ de todas las imágenes, se calcula el promedio para la consulta AVR_q mediante Eq. (2.13), el cual es modificado considerando la cantidad de imágenes en la categoría de consulta R_q, mostrado en Eq. (2.14), Asimismo, se normaliza mediante Eq. (2.15). Finalmente, *ANMRR* resulta del promedio obtenido en todas las consultas realizada como se muestra en Eq. (2.16).

$$Rank_n = \begin{cases} k & si\ k \leq C_q \\ 1.25 \times C_q & en\ otro\ caso \end{cases} \quad (2.11)$$

$$C_q = \min\left(4 \times R_q, 2 \times max\left(R_q, \forall_q\right)\right) \quad (2.12)$$

$$AVR_q = \left|R_q\right|^{-1} \sum_{n=1}^{R_q} Rank_n \quad (2.13)$$

$$MRR_q = AVR_q - 0.5 \times \left(1 + R_q\right) \quad (2.14)$$

$$NMRR_q = \frac{MRR_q}{1.25 \times C_q - 0.5 \times \left(1 + R_q\right)} \quad (2.15)$$

$$ANMRR = |Q|^{-1} \sum_{q=1}^{Q} NMRR_q \quad (2.16)$$

2.2 Antecedentes

Dentro de los trabajos relacionados realizados dentro del Centro nacional de investigación y desarrollo tecnológico CENIDET, se encuentran dos proyectos de maestría que anteceden esta investigación. En 2007 Troncoso [70], desarrolla un sistema de recuperación de imágenes por contenido, que incluye, caracterización, mediante descripciones parciales de los elementos de la imagen, clasificación y recuperación automática de imágenes por contenido. Donde se establece una función hash que clasifica el contenido de las imágenes a partir de sus descriptores para poder indexarlas y realizar la recuperación. El trabajo brinda gran cantidad de información teórica relevante sobre los sistemas de recuperación y descriptores de bajo nivel. Asimismo, Pérez en 2014 [71] implementa un sistema CBIR que emplea los descriptores EHD y CLD propuestos por el estándar MPEG-7 para describir las imágenes. En el cual presenta un conjunto de imágenes desarrollado en las instalaciones de CENIDET, al igual que un software llamado "Recuperación de Imágenes por Ejemplo" (RIPE). El trabajo presenta información y recursos relacionados a los sistemas de recuperación y el estándar MPEG-7. A diferencia de los trabajos que anteceden el proyecto de investigación, lo que se busca en este proyecto es proponer un nuevo descriptor para la recuperación de imágenes naturales en clases semánticas. Lo cual no ha sido abordado en los trabajos anteriores, que se concentran en la creación del sistema, utilizando descriptores clásicos y no en proponer un descriptor de imágenes.

2.3 Estado del arte

Se llevó a cabo un proceso de registro y selección de la literatura científica encontrada en el estado del arte. Se realizaron búsquedas en diferentes fechas entre el 2019 y 2023. La búsqueda se concentró en aquellos trabajos que se consideraron relevantes para el proyecto, es decir aquellos que tenían una mayor similitud al proyecto como los que proponían un descriptor para la recuperación de imágenes naturales o recuperación de imágenes en general. Las búsquedas se realizaron utilizando diferentes ecuaciones de búsqueda, filtros y criterios de ordenamiento de los documentos. Las ecuaciones de búsqueda se seleccionaron a partir de palabras clave encontradas en artículos en el estado del arte con mayor similitud al proyecto, así como de palabras clave consideradas descriptivas para la investigación en cuestión como: *"Image Retrieval, CBIR, Image representation, High level features, Image Descriptor*", así mismo ecuaciones de combinaciones entre ellas como: allintitle: ("Novel" OR "New" OR "efficient" OR "improve" OR "modified")("Descriptor" OR "Representation" OR "Feature")("CBIR" OR "image" OR "Retrieval"). Realizada principalmente en inglés y en español.

Se consideraron como relevantes los documentos obtenidos en cada búsqueda que no se habían obtenido en búsquedas anteriores. Además, se seleccionaron los trabajos de mayor semejanza al proyecto de investigación, centrándose en descriptores o mejoras a descriptores aplicados a los sistemas de recuperación, teniendo en cuenta la procedencia y el año de publicación. Detallando aquellos que presentaron mejores resultados durante la

búsqueda del estado del arte y se lograron obtener o replicar para utilizarlos en este proyecto de investigación.

En general, se encontró que extraer características de nivel alto de los píxeles resulta difícil, ya que estas características representan conceptos semánticamente significativos en la imagen, como actividades que ocurren en la imagen u objetos presentes en ella.

De los tres enfoques listados por [34] para la extracción de características de alto nivel. Se encontraron trabajos que se basan en el enfoque de ontologías [72]–[79]. La ontología de objetos es un vocabulario que se define a partir de diferentes intervalos obtenidos de las características de bajo nivel de la imagen, por ejemplo, del color se podrían obtener las categorías: "verde claro, verde medio, verde oscuro". Esto proporciona una definición cualitativa de conceptos de consulta de alto nivel, por ejemplo, "cielo" se puede definir como una región de "azul claro" (color), 'uniforme' (textura) y 'superior' (ubicación espacial). A pesar de que las ontologías presentan buenos resultados, son un enfoque dirigido principalmente a la anotación automática de imágenes, lo que igual es interesante en la recuperación basada en texto ya que una de las desventajas que presentan estos sistemas, es el trabajo humano que se requiere para la anotación, sin embargo, requieren de un vasto conocimiento del área de aplicación, así mismo, se ve limitado por la cantidad de términos semánticos y reglas utilizadas. Sin embargo, es un enfoque que sigue siendo utilizado para la descripción de imágenes en términos semánticos para la recuperación de imágenes.

Hay dos ramas principales del aprendizaje automático: el supervisado y el no supervisado. En general, los sistemas se basan en la creación de relaciones entre características de bajo nivel y conceptos de alto nivel, lo que explica su aplicación en diversos problemas. La Figura 2.8 ilustra un esquema representativo del proceso seguido por el aprendizaje automático aplicado a los sistemas CBIR. En los artículos [80]–[89], se proponen metodologías para la recuperación de imágenes utilizando técnicas de aprendizaje automático. A pesar de que, este enfoque es prometedor y ha arrojado resultados interesantes, se enfrenta a desafíos como la necesidad de una gran cantidad de datos para el entrenamiento y la limitación en el tipo de imágenes utilizadas en dicho entrenamiento, así como el tamaño de los conjuntos de datos disponibles.

Por último, el tercer enfoque las plantillas semánticas son definidas como un mapa entre el concepto de alto nivel y las características visuales de bajo nivel, se define normalmente como la característica "representativa" de un concepto calculado a partir de una colección de imágenes de muestra [90], [91]. Existen algunas ideas que surgen a partir de las plantillas semánticas como la plantilla visual semántica propuesta en [92], que vincula las características de imagen de bajo nivel con conceptos de alto nivel, es decir, un conjunto de iconos (escenas / objetos) de ejemplo, denotan una vista personalizada de conceptos como reuniones, atardeceres. Algunos trabajos más recientes basados en plantillas semánticas son [93], [94]. Sin embargo, las plantillas semánticas han tenido muy poco interés en los últimos años por lo que no se tiene suficiente investigación para competir con los resultados de otros enfoques.

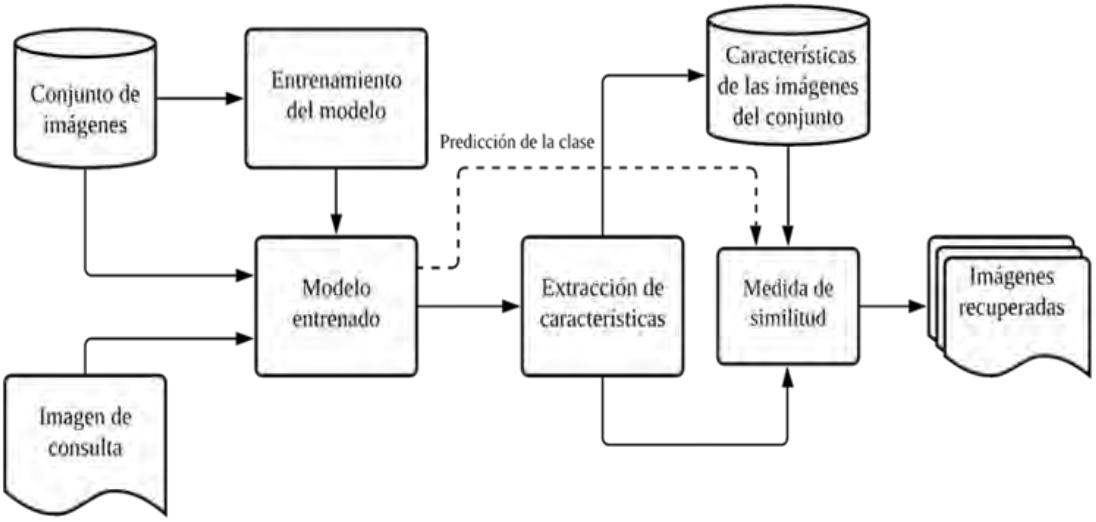


Figura 2.8 Diagrama del framework utilizando en [87]

En la revisión se encontró dos enfoques recientes en el que se centra en la obtención de mejores descriptores visuales, utilizan descriptores derivados a partir de sus características de bajo nivel que consiguen representar semántica de alto nivel. Entre los más comunes están los descriptores que utilizan teorías visuales para generar una descripción partiendo de las relaciones con sus características de bajo nivel. Asimismo, recientemente con el impacto que ha tenido el aprendizaje profundo surgen los descriptores basados en características profundas que aprovechan las características y relaciones generadas mediante el proceso de aprendizaje.

A pesar de que en ocasiones se utilizan algunas técnicas encontradas en otros enfoques o similares, lo que se busca en estos descriptores es obtener una representación de las características de alto nivel, sin la necesidad de asociarlo directamente a un concepto semántico. Mediante la relación de la forma, textura, color, y posición espacial, ya que estas por sí solas parecen ser insuficientes para abordar conceptos semánticos o aspectos más complejos en las imágenes.

Una de las grandes ventajas que presentan los descriptores basados en teorías visuales es que no requieren de entrenamiento, ni conocimiento o términos semánticos asociados. Por lo que el vector obtenido no es asociado o delimitado a un tipo de imágenes en concreto, el costo computacional y humano requerido es menor. A si mismo son descriptores con menor dimensión lo que facilita su almacenamiento e indexación, para una futura estandarización. Lo que se encontró con mayor relevancia a lo que se busca en el proyecto de investigación.

2.3.1 Descriptores que utilizan la integración de características

Entre los trabajos recientes de descriptores que utilizan la integración de características se encuentra [25]. El descriptor propuesto parte de la teoría de la integración de características, la cual es una teoría de la percepción y la atención, que explica cómo un individuo combina

piezas de información observable sobre un objeto para formar una percepción completa. Utilizando esta información integrando características de color en el espacio *HSV* y borde, proponen un descriptor de histograma llamado "*Multi-Integration Features Histogram*" (MIFH). Siguiendo la arquitectura del modelo de integración múltiple de características mostrado en la Figura 2.9.

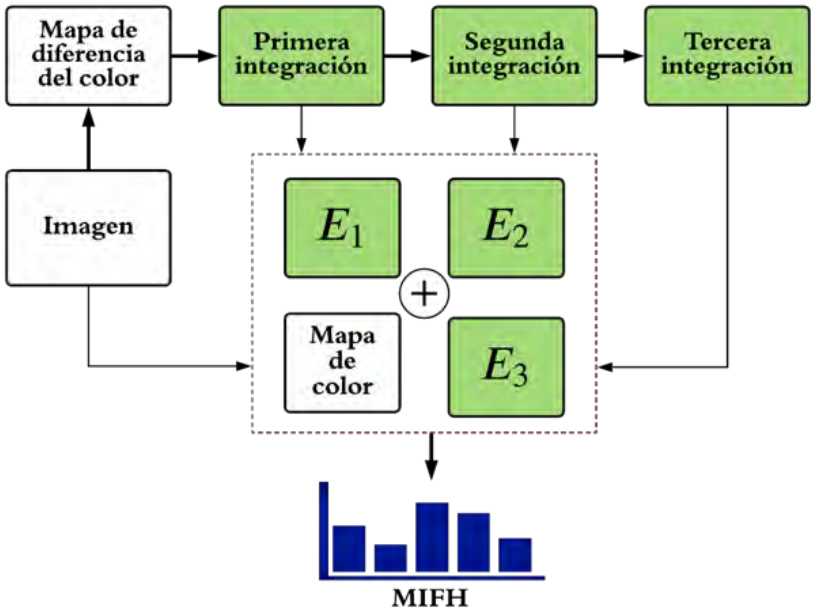


Figura 2.9 Arquitectura del modelo de integración múltiple de características [95]

El descriptor inicia extrayendo las características del color en HSV, y cada canal es cuantificado en 6, 3, y 3 *bins* respectivamente. Dando como resultados $M_0(x, y)$, con un total de 54 combinaciones de color. Asimismo, se obtiene un mapa de diferencia del color, comenzando con la transformación de la imagen HSV de coordenadas cilíndricas a cartesianas. Una vez obtenida la imagen en coordenadas cartesianas, el mapa de diferencia de color es obtenido utilizando la Eq. (2.17). Donde el píxel central está representado por (x_0, y_0) y sus ocho vecinos como $(x_i, y_1), i\epsilon[1,2,3,\ldots,8]$, utilizando $b = 1.0$ para la función sigmoide. Al finalizar, aplica un submuestreo de escala dos sobre $C(x, y)$ obteniendo $S(x, y)$. La primera integración de las características del borde se realiza utilizando Sobel horizontal y vertical en $S(x, y)$, dando como resultado un mapa de borde $G_0(x, y)$. El resto de los mapas se obtienen siguiendo la Eq. (2.18). Donde $\mathcal{g}(.)$ representa el operador Sobel horizontal y vertical. Los cuatro mapas se combinan como se muestra en Eq. (2.19). Asimismo, se aplica un submuestreo de cada uno de los mapas.

$$C(x, y) = sigmoid\left(\sum_{i=1}^{8} \sqrt{(\Delta H_i)^2 + (\Delta S_i)^2 + (\Delta V_i)^2} + b\right) \tag{2.17}$$

$$\Delta H_i = h(x_i, y_i) - h(x_0, y_0),$$

$$\Delta S_i = s(x_i, y_i) - s(x_0, y_0),$$
$$\Delta V_i = v(x_i, y_i) - v(x_0, y_0)$$

$$G_i(x,y) = \begin{cases} g\big(\mathcal{K}_i(x,y)\big), & i \in [1,2,3], \\ donde\ \mathcal{K}_i(x,y) = S(x,y) + \sum_{k=1}^{i} G_{k-1}(x,y) \end{cases} \quad (2.18)$$

$$E_1(x,y) = \sqrt{\sum_{i=0}^{3} \big(G_i(x,y)\big)^2} \quad (2.19)$$

La segunda integración se aplica como en el paso anterior, tomando cada uno de los submuestreos de los mapas anteriores. Dando como resultado 16 mapas $G_{ij}(x,y), i,j \epsilon [0,1,2,3]$. Los mapas se combinan como se muestra en la Eq. (2.20) y Eq. (2.21). Finalmente se aplica un submuestreo a cada $F_i(x,y)$.

$$F_i(x,y) = \sqrt{\sum_{j=0}^{3} \Big(G_{ij}(x,y)\Big)^2}, \quad i \in [0,1,2,3], \quad (2.20)$$

$$E_2(x,y) = \sum_{i=0}^{3} F_i(x,y) \quad (2.21)$$

Para la tercera integración $E_3(x,y)$ se utilizan los submuestreos de cada $F_i(x,y)$ y combinándolos como en Eq. (2.21). A su vez, $E_1(x,y), E_2(x,y)$, y $E_3(x,y)$ se cuantifican en 16 *bins* cada uno obteniendo $M_1(x,y), M_2(x,y)$, y $M_3(x,y)$ respectivamente. Por último, se representa en un histograma la coocurrencia de cada $M_i(\mathrm{x}, \mathrm{y}), \mathrm{i} \in [0, 1, 2, 3]$. Se concatenan los cuatro histogramas y se aplica una transformación logarítmica para ajustar el rango de los valores en el histograma. Asimismo, existen trabajos que integran diferentes características de bajo nivel para mejorar la recuperación de imágenes a partir de su combinación como [96]–[98], que integra características del color, textura y forma. A su vez, se encontraron descriptores que se basan en la integración de descripciones locales como en [99], que utilizando SIFT y SURF generan una descripción para mejorar los resultados en la recuperación de imágenes. Por otro lado, existen propuestas que integran características de bajo nivel y profundas como en [100], al igual que con técnicas de aprendizaje automático [36], algunas otras a partir de su fusión para mejorar la recuperación como en [101]–[105]. Sin embargo, la gran mayoría de descriptores obtiene un vector de gran longitud que dificultaría su indexación y no superan los resultados del descriptor MIFH.

2.3.2 Descriptores que utilizan estructuras

Por otra parte, existen descriptores que utilizan las estructuras como en [106], que se propone un nuevo detector de estructuras locales, simulando el mecanismo de selección de la orientación humana basado en el espacio de color CIELAB. Obtienen un descriptor basado en el detector de estructuras de gradiente, y proponen un sistema de representación discriminativo, "*Gradient Structures Histogram*" (GSH), para describir el contenido de la imagen utilizando colores, orientaciones de los bordes e intensidades como restricciones, y usarlo para CBIR. En [107] proponen una nueva variante basada en "*Multi-Trend Structure Descriptor*" (MTSD) [108], este caracteriza la imagen utilizando la correlación entre las estructuras de color, las orientaciones de los bordes y la textura a nivel local de forma independiente y luego se integran como un único vector de características. La variante que proponen llamada: "*Multi-Direction and Location Distribution of Pixels in Trend Structure*" (MDLDPTS), con algunas diferencias tratando de mejorar la calidad de la recuperación, como la diferencia de codificar una matriz de características de valores cuantificados de color, orientación de borde y textura contra orientaciones de tendencias iguales, pequeñas y grandes.

Algunos descriptores aprovechan la información de los tipos de estructuras para generar la descripción de características, como se propone en [109], donde se presenta el "*Structure Elements' Descriptor"* (SED). El descriptor se basa en estructuras, trabaja en el espacio de color HSV cuantificado a $72\ bins$ y detecta cinco diferentes elementos de estructura. Primero establecen los diferentes elementos de estructuras en $0°, 90°, 45°, 135°,$ y sin dirección, de tamaño 2 × 2. En la segunda etapa las estructuras son detectadas en la imagen cuantificada, obteniendo el "*Structure Elements' Histogram*" (SEH) a lo largo de la imagen como se muestra en la Figura 2.10.

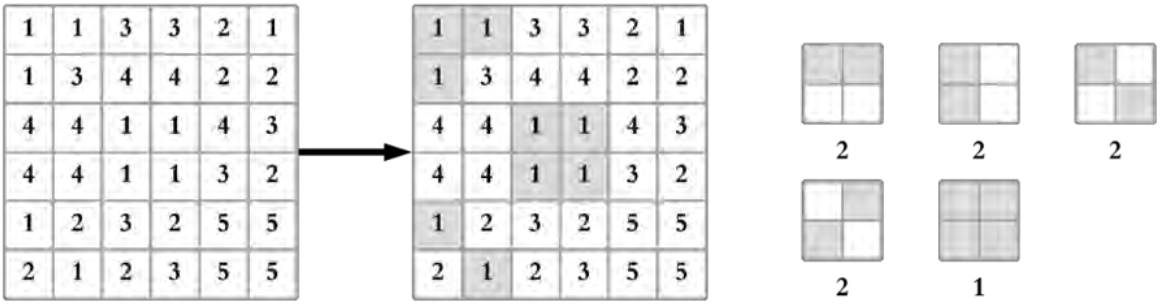


Figura 2.10 Obtención del SEH [109]

El proceso de la segunda etapa se repite en cada uno de los 72 colores dando un vector de 360 valores. Finalmente, los valores son normalizados para hacerlo invariante a escala. Para medir la similitud entre imágenes, los autores proponen su propia medida de similitud mostrada en Eq. (2.22). Donde Q_i, es el valor en la posición i del vector obtenido con la imagen de consulta Q, y T_i el valor en la posición i del vector de una imagen T en el conjunto de imágenes. De los más recientes que ocupan estructuras se encuentran [13], [110]–[114].

$$L_s(T,Q) = \sum_{i=1}^{360} \frac{|Q_i - T_i|}{1 + Q_i + T_i} \quad (2.22)$$

Algunas de los descriptores propuestos en los últimos años se basan en la teoría del "*Texton*", los *Textons* se refieren a estructuras fundamentales en imágenes (y videos), que se consideran los átomos de la percepción visual humana pre atenta [50]. Los *Textons* vienen en forma de manchas atribuidas con color, longitud y orientación (es decir, líneas, manchas alargadas y puntos) [50], [51], [115]. Un ejemplo reciente de estos trabajos es [116], donde presentan un descriptor basado en "*Multi Texton Histogram*" (MTH), llamado: "*Complete Multi-Texton Histogram*" (CMTH), que incorpora información sobre el color, la orientación de los bordes y la distribución del texto dentro de la imagen. El descriptor se evalúa tanto en sistemas de recuperación como clasificación de imágenes. Algunos otros trabajos donde se utilizan los *Textons* son [117]–[120].

2.3.2.1 Descriptores que utilizan microestructuras

Otros descriptores actuales se basan en lo que llaman microestructuras, estas microestructuras se proponen en [121], y parte del pensamiento donde el contenido significativo de las imágenes naturales se compone de muchas microestructuras universales, por tanto, si pudieran extraerse estas microestructuras y describirlas de forma eficaz, podrían servir para la comparación y análisis de diferentes imágenes. Su trabajo se inspira en la teoría de integración de características [49], que se divide en dos etapas; la etapa de preatención, y la etapa de atención. Liu define las microestructuras en [121], como la colección de ciertos colores subyacentes, que tienen una orientación de borde similar o igual en un espacio de color uniforme. Lo más destacado de los colores subyacentes es que pueden combinar señales de color, textura y forma como un todo, a diferencia de la teoría del *textons* de Julesz que se centra en analizar texturas regulares, las microestructuras pueden considerarse como la extensión de los *textons* o la versión en color de *textons*, ya que las microestructuras involucran información de color, textura y forma.

El descriptor de microestructuras "*MicroStructure Descriptor*" (MSD) propuesto por Liu en [121], utiliza tanto el color como la textura. MSD transforma del espacio de color RGB a HSV para detectar las características de la microestructura. Asimismo, se cuantifica la imagen en 72 colores, específicamente los canales H, S y V son cuantificados de manera uniforme con 8, 3, y 3 *bins* respectivamente. Una vez obtenida la imagen en HSV cuantificada, la imagen se transforma de coordenadas cilíndricas a cartesianas para obtener la orientación del borde utilizando Sobel en cada canal. Utilizando los resultados obtenidos en cada canal los autores calculan el vector resultante de las tres direcciones de borde y lo cuantifican en 6 *bins*. Con la información generada obtienen los micro mapas, como se muestran en la Figura 2.11.

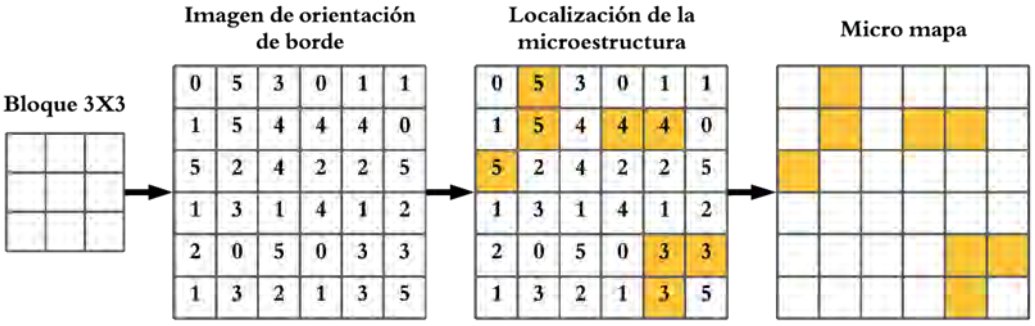


Figura 2.11 Ejemplo de obtención del micro mapa [121]

Los cuatro micro mapas se unen en un mapa de microestructuras obteniendo todas las estructuras encontradas en la imagen, que se utilizan para extraer la información del color en la imagen HSV. Finalmente, se obtiene el histograma normalizado, detectando las ocurrencias en sus ocho vecinos de cada valor en la información extraída del color. Dando como resultado un histograma de 72 valores y utilizando como medida de similitud la distancia de Manhattan Eq. (2.23), donde Q_i, es el valor en la posición i del vector obtenido con la imagen de consulta Q, y T_i el valor en la posición i del vector de una imagen T en el conjunto de imágenes.

$$L_1(Q,T) = \sum_{i}^{N} |Q_i - T_i| \qquad (2.23)$$

Uno de los nuevos descriptores basados en microestructuras es el "*Correlated Microstructure Descriptor*" (CMSD), propuesto en [122], donde mencionan que CMSD representa semántica de alto nivel al identificar microestructuras mediante el establecimiento de correlaciones entre la orientación de la textura, el color y las características de intensidad. CMSD se podría considerar una mejora al descriptor MSD, ya que a diferencia de MSD el CMSD identifica microestructuras mediante el establecimiento de correlaciones entre la orientación de la textura, el color y las características de intensidad. Al igual que el MSD utiliza el espacio de color HSV y lo cuantifica en 72 valores, y las características de intensidad son extraídas del canal V del espacio de color HSV, las cuales se cuantifican en 10 valores. La forma de obtener la dirección de borde varía con respecto a MSD ya que además de utilizar Sobel horizontal y vertical, utiliza Sobel en direcciones diagonales 45° y 135°, una vez obtenida la dirección de borde, se cuantificada en 6 valores. La obtención de los micro mapas se realiza de la misma forma que el MSD, sin embargo, CMSD utiliza dos características para la obtención del mapa de microestructuras, y considera tanto la información de color, como la de los bordes e intensidad, como se muestra en la Figura 2.12.

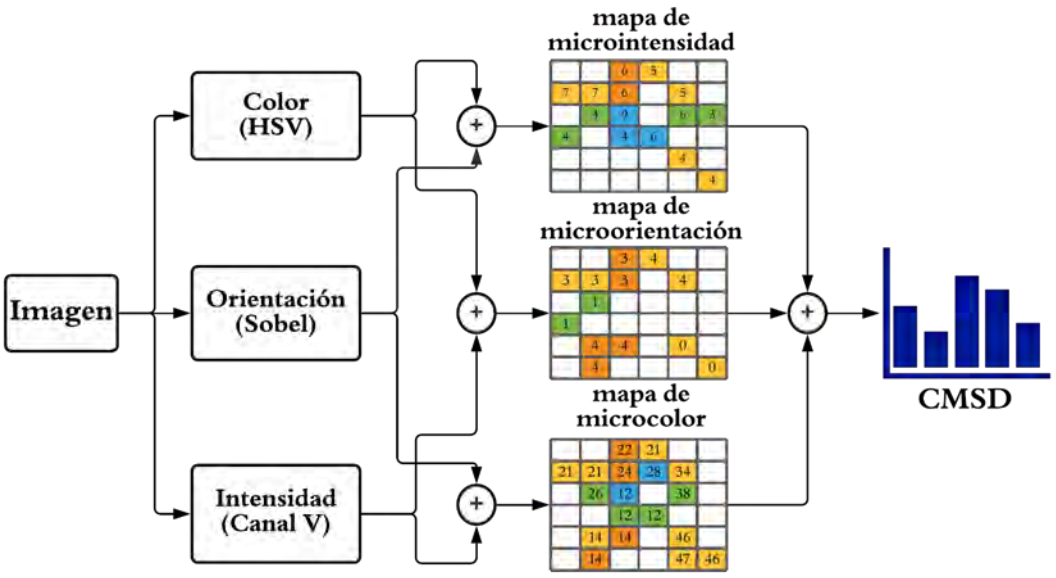


Figura 2.12 Diagrama simplificado de "Correlated Microstructure Descriptor" [122]

Esto da como resultado un descriptor con un vector de 88 valores, que es la concatenación de los tres histogramas obtenidos por cada una de las características $72 + 6 + 10$. CMSD utiliza la distancia *Manhattan* Eq. (2.23), como medida de similitudAdemás, se encontró otro trabajo en el que se presenta un descriptor llamado "*Composite Micro Structure Descriptor*" (ComMSD) [123], que a diferencia del CMSD propuesto por Dawood, combina MTH y MSD.

2.3.3 Descriptores que utilizan características profundas

En el campo del aprendizaje profundo se han presentado sistemas que aprovechan las características generadas y sus relaciones durante el proceso de entrenamiento de los modelos profundos. Existe una colección de trabajos reportados en el estado del arte sobre el uso de características profundas en la recuperación de imágenes por contenido [124]–[127]. Asimismo, en [128] Se desarrolla un nuevo método de recuperación de imágenes en color basado en textura y características profundas, que representa la imagen mediante la combinación de histogramas "*Histogram of Oriented Gradients*" (HOG), momentos invariantes de Hu y características profundas. Este método permite una búsqueda eficiente y precisa de imágenes en grandes conjuntos de datos en color. Otro trabajo es [129] donde se presenta un modelo basado en redes neuronales convolucionales profundas llamado MaxNet para la recuperación de imágenes basada en contenido. El sistema propuesto evita la dependencia de características hechas a mano y extrae características profundas directamente de las imágenes, las cuales se utilizan para recuperar imágenes contextualmente similares de la base de datos. El modelo MaxNet propuesto se construye apilando el módulo Incepción actualizado de manera jerárquica.

En [84], se propone un nuevo método CBIR que utiliza el modelo del grupo VGG basado en transferencia de aprendizaje, el algoritmo genético "*Genetic Algorithm*" (GA) y el clasificador de máquinas de aprendizaje extremo "*Extreme Learning Machines*" (ELM). Este

método permite una búsqueda eficiente y precisa de imágenes en grandes conjuntos de datos mediante la extracción de características profundas y la clasificación mediante el ELM, optimizado mediante GA. Kumar presenta un sistema CBIR en [130], donde se presenta una red neuronal convolucional profunda "*Deep Convolutional Neural Network*" (DCNN) modificada basada en VGG16 (M-VGG16) para la extracción de características de imágenes. La M-VGG16 permite una extracción de características más precisa y eficiente en comparación con la VGG16 original, lo que mejora la precisión de la búsqueda en grandes conjuntos de datos de imágenes. Asimismo, en [52], [55], [131] se presenta una revisión de diferentes sistemas utilizando características profundas. A su vez, existen descriptores que utilizan características profundas y se combinan o se usan en conjunto con las características de bajo nivel como en [132]–[138].

2.4 Discusión

En la literatura se han encontrado trabajos relacionados a la reducción de la brecha semántica. Los trabajos emplean metodologías que obtienen una descripción a partir de la relación que existe en las características de bajo nivel. Esto con la finalidad de representar las características semánticas de la imagen, las cuales difícilmente pueden ser representadas utilizando la información obtenida de manera directa en las características de bajo nivel.

Los resultados encontrados en el estado del arte muestran que los descriptores basados en aprendizaje profundo generan resultados superiores en áreas específicas a las que son entrenados como imágenes médicas, animales, o clases especificas con las encontradas en ImageNET. Sin embargo, no se encontraron descriptores aplicados a imágenes naturales ni una comparación entre el resto de los descriptores basados en teorías visuales. Asimismo, el aprendizaje profundo requiere de un entrenamiento y un costo computacional mayor al resto.

Entre los descriptores basados en teorías visuales como lo son los basados en integración de características, estructuras y microestructuras, se encontraron comparaciones entre ellos, donde se observó un mejor desempeño para los basados en microestructuras y en integración de características. Sin embargo, no existe un estándar en las métricas de evaluación ni en los conjuntos de imágenes utilizados. Incluso cuando los descriptores son propuestos para el mismo objetivo, los autores utilizan las métricas de manera diferente y conjuntos de imágenes distintos, asimismo, no siempre se detallan los experimentos realizados.

En general, se logró observar mejoras en las propuestas más recientes como CMSD y MIFH. Sin embargo, existen descriptores como MSD y SED que siguen siendo ocupados como comparación y tomados como base para nuevas propuestas. Por otra parte, el uso de diferentes conjuntos de imágenes, métricas y la falta de detalle en la información proporcionada, dificultan una evaluación y análisis sobre los descriptores actuales.

Entre las métricas más utilizadas se encontraron las métricas *Precision*, *Recall* y *MAP*. Al igual que variaciones de ellas y en algunas ocasiones la métrica ANMRR. En cuanto a los

conjuntos de imágenes se encontraron algunos conjuntos que destacan, ya sea, por la gran cantidad de trabajos que los ocupan o por la cantidad de imágenes y clases que poseen. De los más utilizados son los obtenidos a partir de *Corel Photo Gallery* [139], el conjunto de datos como lo son Corel-1k [140], Corel-5k [141], y Corel-10k [142]. Asimismo, Caltech-101 [143], Caltech-256 [144], CIFAR-10 [145], CIFAR-100 [146], INRIA Holidays [147], Pascal (VOC2012) [148] y ILSVRC (ImageNet) [149].

En los descriptores basados en características profundas se encontró que la gran mayoría de los trabajos utilizan las características obtenidas mediante arquitecturas como VGG y ResNet, siendo VGG-16 y ResNet-18 las más utilizadas en el estado del arte. Estas arquitecturas son ampliamente utilizadas debido a la abundante cantidad de información y trabajos disponibles. Además, los vectores de características suelen obtenerse mediante los modelos pre entrenados con ImageNet, que cuenta con 1,000 clases variadas. Esto permite obtener características que son más generales y pueden aplicarse en diversas áreas del conocimiento. Por otro lado, el uso de características profundas no siempre implica la eliminación de otros descriptores, ya que existen muchos trabajos donde se complementan las características profundas con otros tipos de descriptores, incluso de bajo nivel.

Teniendo en cuenta los resultados encontrados durante la revisión del estado del arte, los descriptores existentes basados en teorías visuales tienden a tener un rendimiento inferior en comparación con otros enfoques, como aquellos basados en características profundas. Sin embargo, estos descriptores mantienen un vector de menor dimensión, lo que facilita su almacenamiento e indexación para lograr una estandarización, además de no requerir entrenamiento. Por lo que siguen siendo un tema de interés en la comunidad científica en el área de recuperación de imágenes por contenido.

3 Análisis de los descriptores

En este capítulo se presenta el análisis realizado a los descriptores del estándar MPEG-7, así como a los descriptores encontrados en el estado del arte. El objetivo del análisis fue obtener una comparación entre los descriptores, utilizando los mismos conjuntos de datos accesibles y comúnmente utilizados en el estado del arte, al igual que las métricas frecuentemente empleadas, bajo las mismas condiciones.

3.1 Implementación

Para la detección de debilidades se utilizaron siete descriptores, dos de los cuales pertenecen al estándar MPEG-7, y cinco descriptores propuestos en el estado del arte. Se tomaron los dos descriptores más utilizados en la recuperación de imágenes del estándar MPEG-7, un descriptor de textura EHD y un descriptor de color CLD. Asimismo, con base en los resultados y fecha de publicación, se implementaron cinco descriptores que utilizan la relación de las características de bajo nivel para mejorar la recuperación de imágenes naturales. La Tabla 3.1 muestra los descriptores extraídos del estado del arte, proporcionando información sobre el tipo de características de bajo nivel empleadas, el método utilizado para obtener información a partir de la relación entre estas características, la longitud del vector resultante, los conjuntos de datos utilizados para su evaluación, la cantidad de imágenes consideradas para la recuperación K y el correspondiente porcentaje de mejora. El porcentaje de mejora se calcula en relación con el descriptor mejor evaluado con el que se comparan, y se evalúa en el conjunto de imágenes donde el descriptor propuesto obtiene su mejor rendimiento.

Considerando que los trabajos se presentan bajo diferentes condiciones ya que toman imágenes aleatorias y consideran diferente número de imágenes recuperadas K para la evaluación, se decidió implementar cada uno de ellos con la finalidad de obtener una evaluación justa y detectar sus debilidades. Los cuales fueron codificados a partir del modelo propuesto por sus autores. Los descriptores CMSD y CMTH fueron proporcionados por sus autores, sin embargo, el descriptor CMSD a lo largo de los experimentos se

encontró con posibles problemas de implementación por lo que no seguía el modelo propuesto por el autor, lo que a su vez afectaba su rendimiento.

Tabla 3.1 Descriptores implementados

Descriptor	Características utilizadas	Método para extraer características de mayor nivel	Longitud del vector	Conjuntos de imágenes	K considerado	Mejora
MSD	• Color • Borde	Microestructuras	72	• Corel-10k • Corel-5k	12	6%
SED	• Color	Estructuras	360	• Corel-1k • Corel-10k	10	2%
CMSD	• Color • Borde • Intensidad	Microestructuras	88	• Corel-1k • Corel-5k • Corel-10k	12	6%
CMTH	• Color • Borde	Texton	283	• Corel-10k	-	-
MIFH	• Color • Borde	Integración de características	112	• GHIM-10k • Corel-10k • Corel-5k	-	4%

Los experimentos se realizaron en Windows 10, con el software MATLAB en su versión R2020a. Se utilizaron los conjuntos de imágenes: Corel-1k; Corel-5k; y caltech-101, del cual, se tomaron solo 100 de las 101 clases ya que no se consideraron las imágenes en escala de grises. Se tomaron 10 imágenes aleatorias por clase como imagen de consulta, dando un total de 100 consultas por descriptor, consiguiendo así un total de 700 consultas

con Corel-1k; 3,500 con Corel-5k; y 7,000 con Caltech-101. Por fines prácticos y considerando que las personas buscan tener un resultado en las primeras imágenes recuperadas se estableció el sistema en $K = 12$, es decir las primeras 12 imágenes más relevantes de cada consulta. Se evaluaron los siete descriptores, considerando la mejor versión o combinación, ya que, tanto el CMSD como el MIFH presentaron algunos inconvenientes los cuales se explican en detalle a continuación.

3.1.1 Ajustes y corrección de los descriptores

Analizando el código del descriptor CMSD proporcionado por el autor, se detectó un total de cuatro problemas: la omisión o unión de valores; el uso equivocado de variables; así como, el uso incorrecto de la función $seno$ y $coseno$ de Matlab; y finalmente la conversión de coordenadas cilíndricas a cartesianas. Los problemas provocaban valores nulos, ruidosos o erróneos en la representación del descriptor, por lo que se decidió modificar el código de forma que siguiera fielmente el modelo que propone.

Se compararon los resultados obtenidos con los ajustes realizados al descriptor. Los resultados de la evaluación utilizando los tres conjuntos de imágenes y usando las cuatro métricas, se muestra en Tabla 3.2. Donde la versión corregida se establece como CMSD-2, y la versión proporcionada por el autor como CMSD-1. Los resultados obtenidos con la evaluación mostraron que el descriptor corrigiendo las discrepancias entre el modelo y el algoritmo, obtuvo mejores resultados en promedio. Por lo que se decidió utilizar esta versión del algoritmo, para los experimentos posteriores.

Tabla 3.2 Evaluación del descriptor CMSD

	Corel-1k		**Corel-5k**		**Caltech-101**	
	CMSD-2	CMSD-1	CMSD-2	CMSD-1	CMSD-2	CMSD-1
P	77.33%	75.67%	34.10%	32.48%	10.90%	10.16%
R	9.28%	9.08%	4.09%	3.90%	1.71%	1.56%
MAP	8.64%	8.40%	3.05%	2.90%	1.09%	0.99%
ANMRR	1.31%	1.41%	3.90%	4.00%	4.54%	4.58%

Con respecto al descriptor MIFH, en la literatura no se establecía el incremento en x, ni el incremento en y para la obtención de la coocurrencia ni el tipo de submuestreo realizado. Por lo que se realizó una evaluación con diferentes combinaciones, usando tres submuestreos diferentes, promedio (AVG); mínimo (MIN); y máximo (MAX), así como, tres combinaciones de incrementos: $x = 1,\ y = 0$; $x = 0,\ y = 1$; y $x = y = 1$. Obteniendo así, $3 * 3 = 9$ combinaciones. Los resultados se muestran en Tabla 3.3 utilizando el conjunto de imágenes corel-1k y en la Tabla 3.4 con el conjunto de imágenes Corel-5K. Donde las variantes se denotan de acuerdo con el tipo de submuestreo y los incrementos utilizados, en la forma "submuestreo$-xy$". De la evaluación, se tomó la combinación que se encontró con mejores resultados, en este caso la combinación AVG-01, ya que obtuvo mejores resultados en la evaluación con Corel-1k y en Corel-5k.

Tabla 3.3 Evaluación combinaciones MIFH con Corel-1k

	P	R	MAP	ANMRR
AVG-11	73.58%	8.83%	7.93%	1.54%
AVG-10	72.58%	8.71%	7.95%	1.59%
AVG-01	75.00%	9.00%	8.21%	1.45%
MAX-11	73.50%	8.82%	8.06%	1.54%
MAX-10	74.33%	8.92%	8.15%	1.49%
MAX-01	73.75%	8.85%	8.07%	1.53%
MIN-11	74.08%	8.89%	7.97%	1.51%
MIN-10	73.58%	8.83%	7.94%	1.54%
MIN-01	74.00%	8.88%	8.04%	1.51%

Tabla 3.4 Evaluación combinaciones MIFH con Corel-5k

	P	R	MAP	ANMRR
AVG-11	28.88%	3.46%	2.40%	4.21%
AVG-10	28.72%	3.44%	2.40%	4.22%
AVG-01	28.78%	3.45%	2.43%	4.22%
MAX-11	27.82%	3.33%	2.30%	4.28%
MAX-10	28.43%	3.40%	2.35%	4.24%
MAX-01	28.55%	3.42%	2.37%	4.24%
MIN-11	28.25%	3.38%	2.37%	4.25%
MIN-10	28.22%	3.38%	2.34%	4.25%
MIN-01	28.75%	3.44%	2.41%	4.22%

3.2 Detección de debilidades

Para la detección de debilidades se realizó una evaluación de los siete descriptores, con las configuraciones antes mencionadas. Los resultados se muestran en Tabla 3.5, Tabla 3.6 y Tabla 3.7. En los resultados se encontró que los descriptores MIFH y CMSD obtienen las mejores evaluaciones en los conjuntos de imágenes Corel-1k y Corel-5k; sin embargo, en el conjunto de imágenes Caltech-101, el descriptor EHD obtiene un mejor desempeño que el resto de los descriptores, seguido del descriptor CLD. El resultado podría deberse a que los descriptores propuestos por el estándar están orientados principalmente a recuperaciones en nivel uno y dos es decir de formas, colores y objetos a diferencia del resto de descriptores, que son orientados a recuperaciones más complejas. Aunado a esto, Caltech-101 contiene clases no balanceadas, por lo que la métrica más fiable es ANMRR en la cual la diferencia es menor que el resto de las métricas.

Tabla 3.5 Resultados obtenidos con Corel-1k

	CLD	EHD	MSD	CMSD	SED	MIFH	CMTH
P	49.00%	56.58%	70.75%	77.33%	61.17%	75.00%	48.50%
R	5.88%	6.79%	8.49%	9.28%	7.34%	9.00%	5.82%
MAP	4.92%	5.70%	7.58%	8.64%	6.28%	8.21%	4.65%
ANMRR	3.02%	2.56%	1.70%	1.31%	2.27%	1.45%	3.04%

Tabla 3.6 Resultados obtenidos con Corel-5k

	CLD	EHD	MSD	CMSD	SED	MIFH	CMTH
P	10.22%	10.33%	26.50%	34.10%	26.80%	28.78%	10.42%
R	1.23%	1.24%	3.18%	4.09%	3.22%	3.45%	1.25%
MAP	0.64%	0.68%	2.20%	3.05%	2.26%	2.43%	0.64%
ANMRR	5.37%	5.37%	4.37%	3.90%	4.35%	4.22%	5.36%

Tabla 3.7 Resultados obtenidos con Caltech-101

	CLD	EHD	MSD	CMSD	SED	MIFH	CMTH
P	14.26%	20.43%	8.87%	10.90%	7.50%	10.51%	7.26%
R	2.18%	3.57%	1.35%	1.71%	1.17%	1.58%	1.16%
MAP	1.59%	2.63%	0.87%	1.09%	0.69%	1.03%	0.69%
ANMRR	4.43%	4.10%	4.63%	4.54%	4.67%	4.57%	4.68%

Se encontró que los descriptores que utilizan más de una característica obtienen mejores resultados con los conjuntos que poseen clases complejas, a excepción del descriptor CMTH que no obtiene resultados superiores a los descriptores clásicos. En general, las clases más complicadas para la mayoría de los descriptores son aquellas referentes a un lugar específico, ya que contenían una gran diversidad de contenidos sin una aparente relación visual, al igual que las clases con contenidos similares a otra, como la clase aviones y la clase objetos voladores del Corel-5k.

3.2.1 Imágenes y clases problemáticas

En la evaluación realizada se encontraron imágenes y clases problemáticas. Algunas de las recuperaciones se detectaron con errores aparentemente sin relación alguna con la imagen de consulta como las imágenes 1 y 6 de la consulta mostrada en Figura 3.1. Para analizar los resultados se optó por seleccionar los descriptores con mejor desempeño en la evaluación, en este caso los descriptores CMSD y MIFH. Se codificó un sistema de recuperación que permitiera seleccionar y cortar partes de una imagen de consulta, con la finalidad de buscar una relación existente entre una zona o parte de la imagen y las imágenes erróneas, intentando detectar el motivo de confusión y la debilidad de los descriptores. Algunos ejemplos de consultas se muestran en Figura 3.2 y Figura 3.3, donde se presentan las imágenes recuperadas con dos cortes de la imagen de consulta de la Figura 3.1, colocando como nombre las coordenadas del corte.

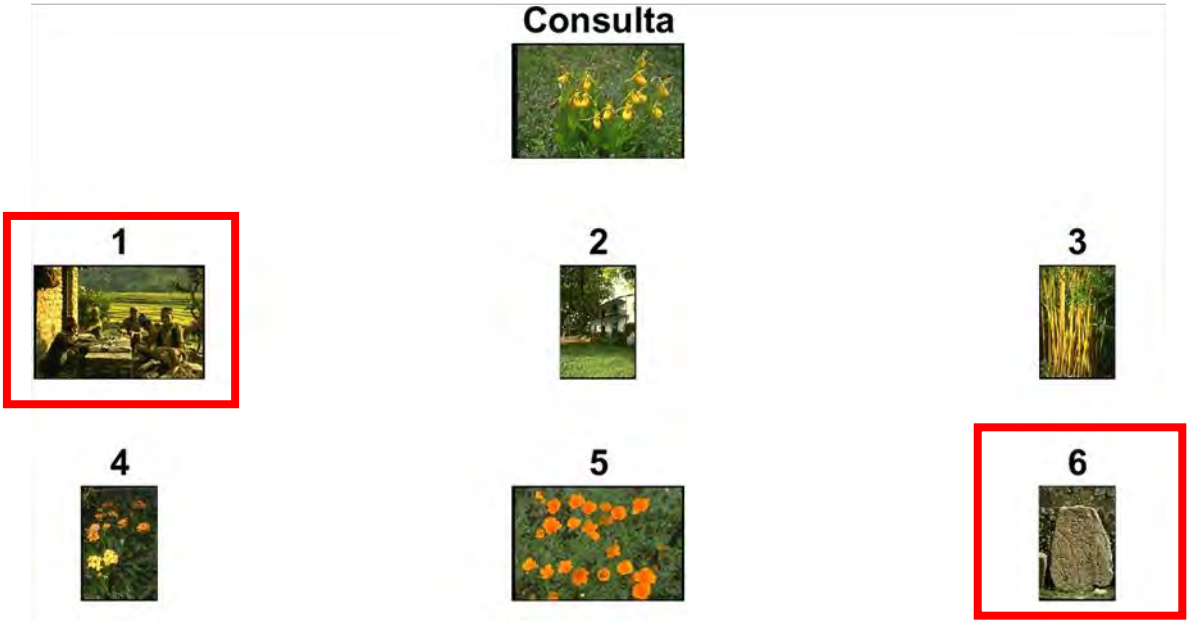


Figura 3.1 Recuperación de flores con CMSD

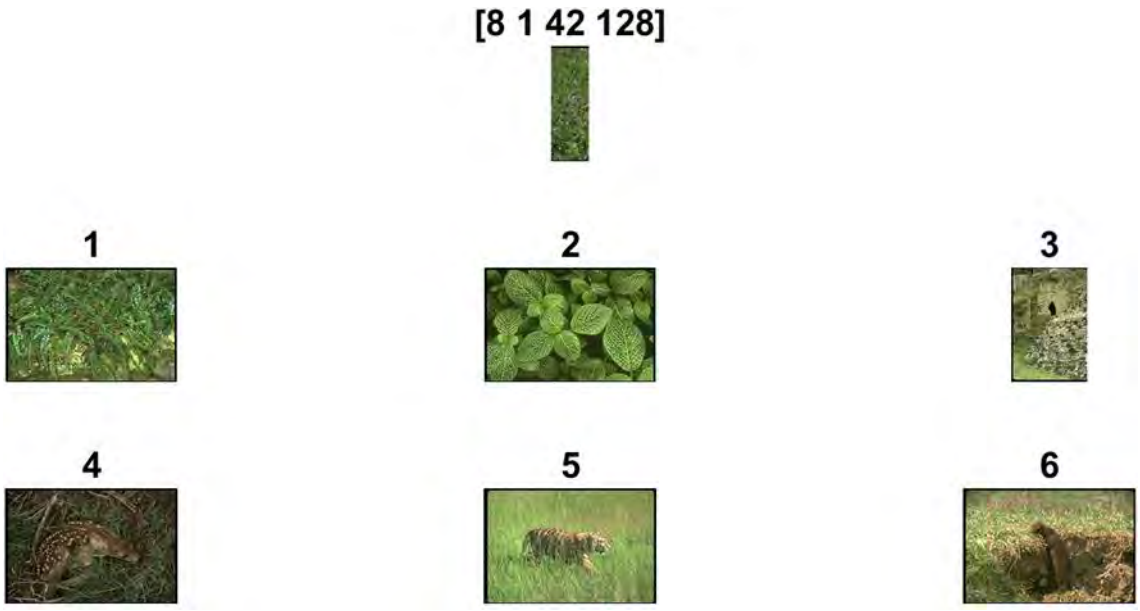


Figura 3.2 Recuperación utilizando corte del fondo de la imagen de consulta

A lo largo del experimento se encontró, que, en algunos de los casos, la explicación de la confusión podría deberse a la información introducida por zonas de la imagen, que pueden encontrase en el fondo, ya que, en la consulta realizada, dando como entrada un corte de la zona del fondo, se recuperaba la imagen errónea en una mejor posición, incluso en algunas situaciones, se recuperaban otras imágenes parecidas. Sin embargo, existen casos en los que no se logró encontrar una relación entre las diferentes zonas de la imagen y la

imagen errónea, por lo que aparentemente el error en la recuperación no se relaciona con una región especifica dentro de la imagen, sino a la información en su conjunto.

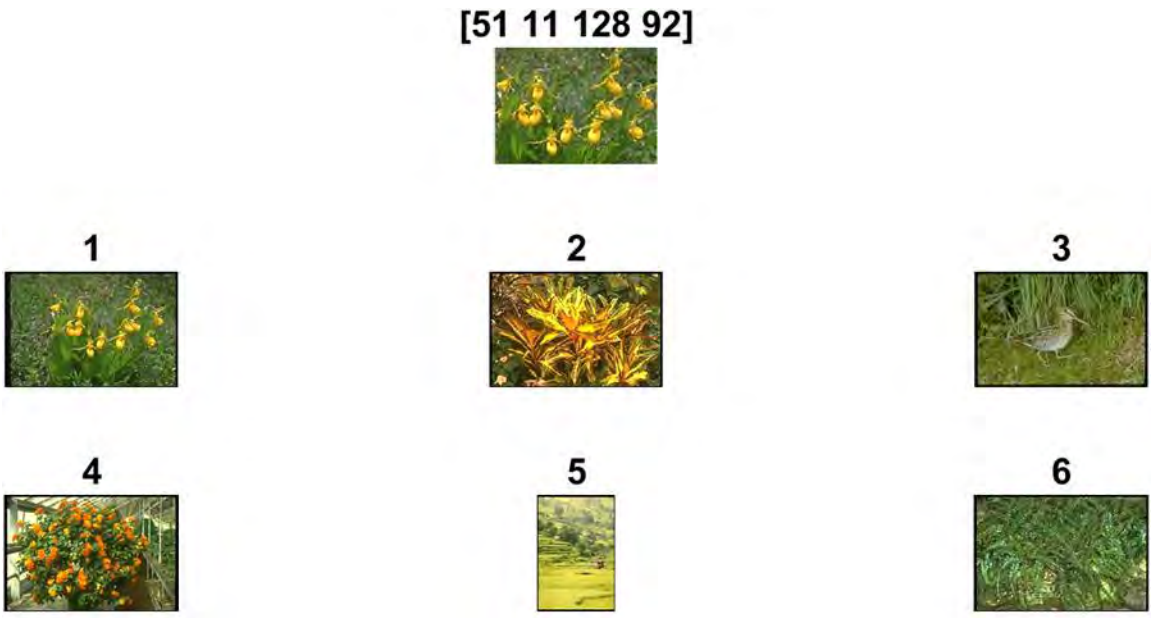


Figura 3.3 Recuperación utilizando corte centrado de la imagen de consulta

Asimismo, se encontró que las imágenes podrían tener una relación en la dirección o proporción de las texturas y color, a pesar de no estar distribuidas de la misma manera. Por ejemplo Figura 3.4, a pesar de no tener una relación visual aparente, las personas y el autobús tienen una perspectiva similar, y contienen colores similares, al momento de cuantificarlos. Asimismo, se observó que las consultas en imágenes saturadas o con una baja variedad de colores, es decir con un histograma de color aglomerado en un rango, recuperaban imágenes saturadas, incluso a pesar de no tratarse aparentemente de los mismos colores, como el error obtenido en Figura 3.5.

Con el propósito de identificar las posibles debilidades responsables de los errores en la recuperación, se procedió a inducir deliberadamente el error seleccionando imágenes que cumplieran con características que pudieran propiciar una recuperación incorrecta. Luego, se llevó a cabo un análisis de estas imágenes identificadas como problemáticas, generando sus histogramas de diferencias, tal como se muestra en la Figura 3.6. Este análisis se enfocó en examinar cada uno de los valores en el vector de características, con la intención de identificar aquellos valores que podrían dar lugar a la confusión o, alternativamente, detectar valores que pudieran utilizarse para prevenirla.

En el experimento se encontró que los valores que causaban confusión a menudo eran los valores en cero del histograma de color. Es decir, en el caso del CMSD las imágenes saturadas se parecían en el histograma de color ya que lo único que los diferenciaba de los 72 valores de color eran de 6 a 10 valores teniendo un total de 62 a 66 valores iguales ya

que no se tenía colores en ese rango. En el caso del descriptor MIFH, a pesar de no tener más que 54 valores de color se presenta el mismo caso ya que al generalizar más el color la diferencia es menor.

Figura 3.4 Recuperación de playa con CMSD

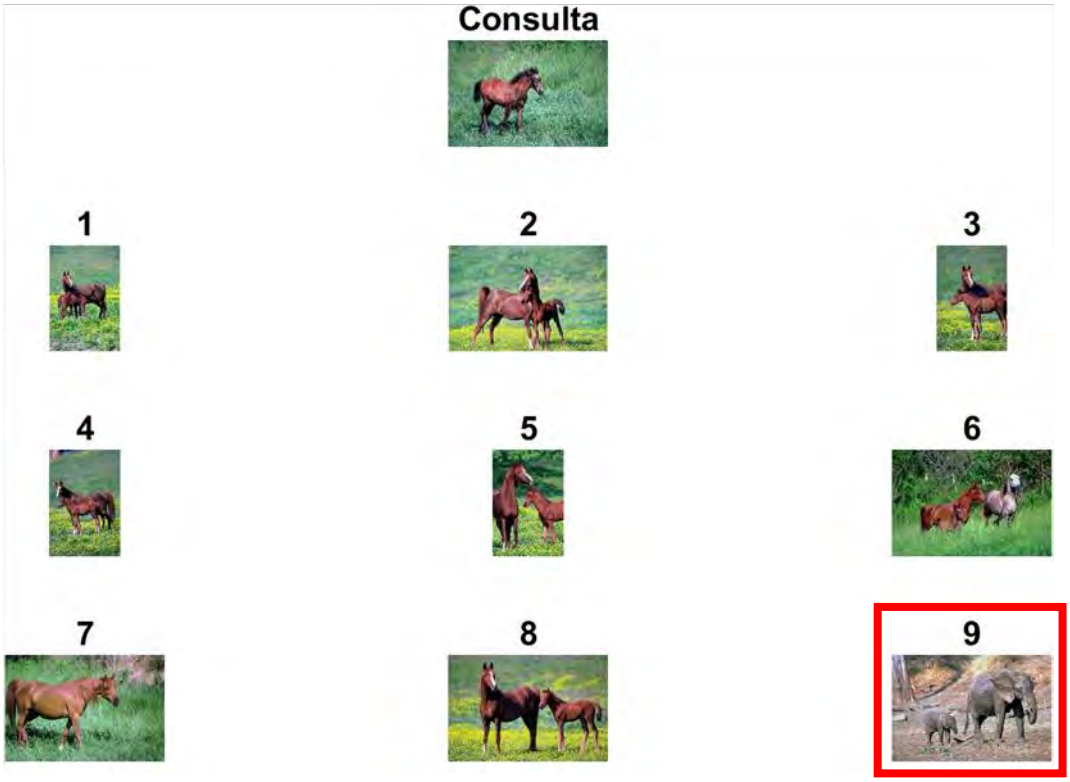


Figura 3.5 Recuperación de caballo con MIFH

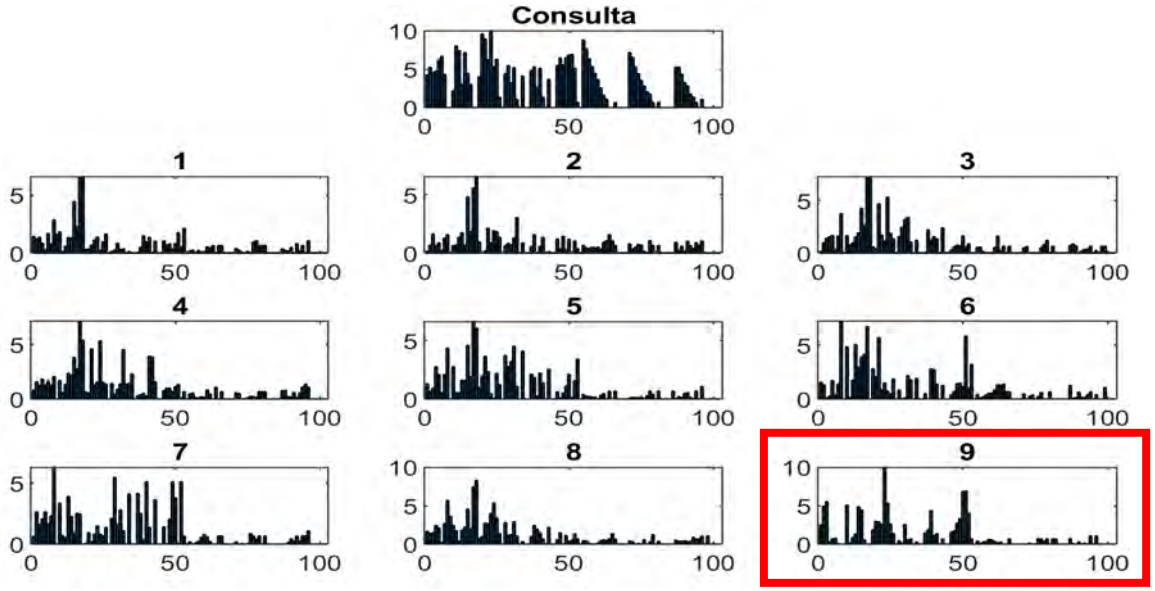


Figura 3.6 Histograma de diferencias de la consulta caballo con MIFH

A su vez, se optó por calcular las medidas de dispersión de los vectores de características. Con la finalidad de encontrar los valores con mayor o menor impacto en la discriminación, del vector de características. Se utilizó el Rango, Eq. (3.1), la desviación estándar, Eq. (3.2) y el coeficiente de variación, Eq.(3.3). Donde X es la población; x cada uno de sus elementos; N, la cantidad de datos; y μ el promedio. Con las medidas de dispersión se buscaron valores con rangos amplios, una desviación estándar pequeña entre las imágenes de la clase y grande entre clases, y con un coeficiente de variación mayor al 26%. Para detectar los valores que tuvieran las mayores diferencias y a la inversa para detectar valores con diferencias nulas o inferiores.

$$Rango = (\max(X) - \min(X)) \tag{3.1}$$

$$\sigma = \sqrt{\frac{\sum(x-\mu)^2}{N}} \tag{3.2}$$

$$CV = \frac{\sigma}{\mu} \tag{3.3}$$

Durante el transcurso de los experimentos, se identificaron las problemáticas previamente aludidas en relación con el descriptor CMSD, debido a la presencia de valores nulos en su vector característico. Asimismo, se constató que los valores en general, asociados al descriptor, no mostraban una consistencia notable en términos de su capacidad discriminativa entre las diversas clases. A excepción de los valores nulos, no se pudo

discernir algún valor específico que presentara un grado significativamente bajo o alto de capacidad discriminativa en comparación con otros.

En lo que concierne al descriptor MIFH, se observó que los valores que representan los bordes de la imagen, específicamente en el rango del 55 al 102, parecen carecer de capacidad discriminativa apreciable, e incluso podrían estar generando ambigüedad en la clasificación. Este fenómeno se hacía especialmente notorio en los cinco últimos valores de cada conjunto de características de borde, es decir, los valores $[66, 67, \ldots, 70, 81, 82, \ldots, 86, 98, 99, \ldots, 102]$. En la gran mayoría de las imágenes, estos valores se mantenían en cero, mientras que, en algunos casos, se observaban valores constantes sin que se pudiera identificar una relación aparente con las clases.

3.2.2 Texturas y color

Buscando las debilidades relacionadas a la textura y el color, así como, sustentar algunas de las debilidades encontradas, se realizaron experimentos utilizando texturas con diferentes colores. Para el experimento se ocuparon imágenes de dos conjuntos de imágenes de Kaggle: *texture dataset* [150]; y *texture* [151]. Se agregó color, transformando la imagen a RGB y saturándola en un canal sin perder el patrón de textura. Con la finalidad de aumentar el número de imágenes y cumplir con las características necesarias para comprobar la confusión de color y textura. Se utilizaron cinco valores diferentes para H $[0°, 45°, 90°, 130°, 180°]$; tres para S $[1, 0.25, 0.17]$; y tres para V $[0.23, 0.5, 0.69]$, todos tomados de forma aleatoria. Un ejemplo de recuperación con las imágenes de textura se muestra en Figura 3.7.

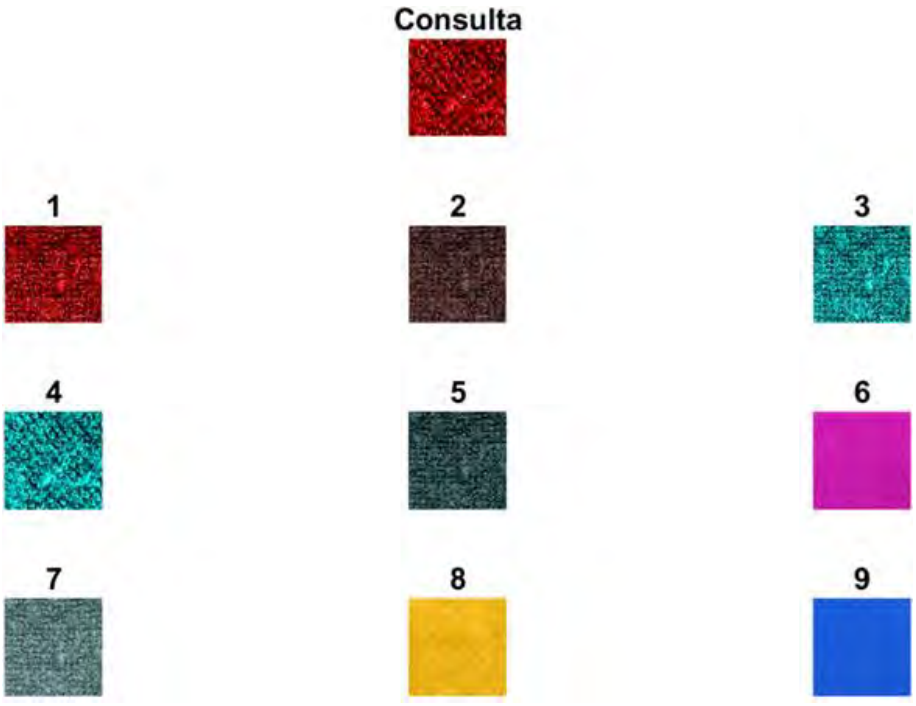


Figura 3.7 Recuperación de texturas con CMSD

En los resultados se presentaron recuperaciones erróneas, donde la textura de la imagen no era tan relevante como el color. Ya que las imágenes recuperadas se parecían en gran medida en los primeros 72 valores en el caso del CMSD, de igual forma en los diez valores del vector de luminosidad que está compuesto principalmente del canal V, por lo que a pesar de presentar un gran parecido en el vector de borde no se consideraban similares ya que solo eran similares en seis valores. A pesar de no ser considerados completamente erróneos ya que las imágenes tienen características muy similares en el tono, la saturación o el valor, se podría considerar que el descriptor tiende a recuperar imágenes con más similitud en el color.

3.3 Discusión

Algunos de los errores se presentan por problemas en los conjuntos de imágenes, ya que existen clases muy similares, o que poseen contenidos muy variados, como es el caso de las clases que pertenecen a una región en específico, por ejemplo, la clase *Roma*, asimismo, existen imágenes que bien podrían pertenecer a más de una clase, o que contienen más información de otra clase que en la que están etiquetadas. Los descriptores clásicos del estándar se desempeñan mejor que los otros en el conjunto de imágenes Caltech-101, por lo que aparentemente al intentar representar características de alto nivel de la imagen, disminuyen su efectividad en consultas de menor nivel de recuperación.

Los descriptores MIFH y CMSD otorgan una gran importancia al color. En el caso del descriptor CMSD, el color representa un 82% del vector, mientras que, en el MIFH, a pesar de ser un porcentaje menor a 53%, el histograma de color se obtiene directamente del espacio de color HSV, lo que significa que esos 54 valores dependen completamente del color de la imagen. Por otro lado, la mayoría de los descriptores utilizan Sobel para la detección de bordes. Aunque en descriptores como CMSD se utiliza una variación de Sobel para abarcar los tres canales de color, existe la posibilidad de perder detalles de los bordes debido a la cuantificación o a la técnica utilizada. Asimismo, es posible que la generalización del color provoca que, colores aparentemente diferentes se consideren iguales, ya que de 2^{24} pasan a 72 y 54 colores.

Algunos valores de los descriptores parecieran ser poco representativos lo que podría estar provocando recuperaciones erróneas, y a pesar de considerar la relación entre dos características de la imagen, se siguen obteniendo resultados erróneos muy parecidos a los de un descriptor de bajo nivel, ya que tienden a darle mayor importancia a una característica, por lo que faltaría analizar y proponer nuevas formas de relacionar la información. A su vez se encontró que en los descriptores de microestructuras únicamente contemplan la información de la relación de las características y no contempla la información del tipo de estructuras lo cual podría ayudar a obtener una mejor discriminación con un menor costo computacional, considerando que es información que ya se ha obtenido.

Finalmente, se han identificado las principales debilidades en los descriptores, que se centran en la gestión del color, abordando tanto su representación como su relevancia en el proceso de análisis. Se han detectado deficiencias en la información de fondo, así como

en la representación misma, ya que el vector contiene valores poco representativos. Estas deficiencias pueden estar relacionadas con las técnicas empleadas para la extracción de características de bajo nivel o con la omisión de información relevante que podría contribuir a una representación más precisa de la imagen.

Para abordar estas debilidades potenciales, proponemos considerar cambios en la cuantificación o en el modelo de espacio de color utilizado. Además, se sugiere la exploración de nuevas métricas de similitud para garantizar que ninguna característica o valor reciba un peso desproporcionado. Asimismo, la aplicación de modelos de atención visual o técnicas gaussianas podría atenuar el impacto de la información de fondo.

En aras de mejorar la representación, recomendamos evaluar diversas metodologías de extracción de características de bajo nivel y considerar otras características que puedan enriquecer la descripción de la imagen. También es crucial perfeccionar las técnicas empleadas para la correlación de características y ajustar los rangos de cuantificación de cada característica, dado que su comportamiento puede variar significativamente debido a sus distintos niveles de dispersión.

4 Propuestas para mejorar la representación de características de alto nivel

En el capítulo se presentan las propuestas de mejora realizadas considerando las debilidades detectadas. Se optó por proponer diferentes modificaciones a los dos descriptores mejor evaluados en la detección de debilidades, siendo los descriptores CMSD y MIFH, considerando como base su mejor ajuste o versión.

4.1 Propuestas basadas en integración de características

El descriptor MIFH está basado en la teoría de integración de características, que puede considerarse como una teoría de la percepción, utiliza el espacio de color HSV y las características del borde para la representación de las imágenes.

4.1.1 Oportunidades de mejora en Multi-Integration Feature Histogram

El descriptor MIFH entrega un vector de 102 características, que podrían dividirse en dos secciones, de la 1 a la 54, para el color y de la 55 a la 102 para la integración de características del borde. En general el descriptor obtiene buenos resultados en la recuperación, sin embargo, en los valores que representan la integración del borde, se han encontrado algunos que podrían estar causando confusión. Ya que, en la gran mayoría de imágenes se obtiene un cero, y en algunos casos se presentan valores fijos sin variaciones y sin relación aparente entre clases, principalmente en los últimos 5 valores de cada una de las integraciones de características del borde, es decir, del 66 al 70, del 81 al 86, y del 97 al 102.

4.1.2 Multi-Integration Feature Histogram Reduced

Con la finalidad de mejorar el desempeño del descriptor y considerando que presentó valores nulos o poco representativos en la descripción, se propusieron dos ajustes para reducir el ruido introducido por los valores nulos o fijos sin una relación aparente. El primer ajuste propuesto (MIFH-R1) omite los últimos cinco valores de cada integración de borde.

Para el segundo ajuste (MIFH-R2) se consideran todos los valores, sin embargo, lo últimos cinco valores son agrupados en un único valor.

- MIFH-R1 sigue el mismo enfoque que se propone para MIFH. Sin embargo, se distingue en el proceso de generación de características después de la obtención de las tres integraciones de borde cuantificadas en 16, es decir, $M_1(x,y), M_2(x,y)$, y $M_3(x,y)$. En este punto, se procede a construir un histograma que representa la coocurrencia de cada $M_i\,(\mathrm{x},\,\mathrm{y}),\ \mathrm{i} \in [1,\,2,\,3]$, limitándose a considerar solamente los primeros 11 bins de cuantificación, es decir, los valores en el rango de [0-10].

  Una vez que se han generado los tres histogramas de borde, cada uno con una longitud de 11 *bins*, se concatenan para formar un único vector de características. A este vector se le añaden los 54 valores correspondientes al histograma de color. Posteriormente, se aplica la transformación logarítmica para ajustar el rango de los valores contenidos en el histograma tal como lo hace MIFH.

  El resultado final es un vector de características con una dimensión total de $11 + 11 + 11 + 54 = 87$ valores, que combina las características de borde considerando únicamente los primeros 11 valores de cada integración y las características de color.

- MIFH-R2 tiene como objetivo mitigar la presencia de valores nulos o fijos en las características de borde, sin eliminar información potencialmente relevante para la recuperación de datos. Este ajuste se logra agrupando los últimos cinco valores de cada integración de bordes en un solo valor representativo.

  En este proceso de ajuste, una vez que se han cuantificado las tres integraciones de bordes, se procede a construir un histograma de coocurrencia, en el que los *bins* correspondientes a $[12 - 16]$, se agrupan en un solo *bin*. Esto resulta en la generación de tres histogramas con una longitud de 12 bins cada uno. Adicionalmente, se aplica la transformación logarítmica para normalizar el rango de valores en cada uno de los tres histogramas, incluyendo también el histograma de color. Lo que conduce a la obtención de un vector de características con una dimensión total de $12 + 12 + 12 + 54 = 90$ valores. Este vector combina las características de borde agrupadas y las características de color, lo que facilita la representación de información importante para la tarea de recuperación de datos.

4.1.3 Multi-Integration Feature Histogram Quantification

Considerando que los valores nulos, podrían estar relacionados con la pérdida de información durante la cuantificación, ya sea por los rangos o formas de cuantificar los valores de la integración del borde. Se realizaron dos modificaciones enfocadas en la cuantificación de los valores. En la primera modificación (MIFH-Qm) se establece una cuantificación sobre los mapas de borde utilizando una cantidad menor de *bins* para cada uno y no contempla la tercera integración de borde. Por otra parte, la segunda propuesta (MIFH-Qr) considera diferentes rangos para la cuantificación de las integraciones del borde.

- MIFH-Qm presenta una variante respecto a MIFH, en la que se realiza la cuantificación no en la etapa de integración global, sino en cada mapa de bordes individual $G_i(x, y)$. Por lo que la integración global de bordes no se tiene en cuenta en este enfoque. Para comprender mejor el proceso, se describen los detalles de cuantificación de los mapas de bordes generados en las distintas etapas:

    1. Los valores de cada uno de los cuatro mapas de bordes generados en la primera integración se cuantifican utilizando 2 *bins* para cada mapa. Esto resulta en un total de $2 \times 2 \times 2 \times 2 = 16$ *bins* para los primeros cuatro mapas, para la segunda integración.
    2. En la segunda etapa, se genera la tercera integración de bordes a partir del submuestreo de los mapas de bordes de la segunda integración. En esta fase, se toman los mapas de integración de bordes de la segunda integración, se les aplica el submuestreo y se cuantifican utilizando 8 *bins* cada uno. Esto implica la generación de 16 mapas de borde, denotados como $G_{ij}(x, y), i, j\epsilon[0,1,2,3]$, y luego se integran para obtener cada $F_i(x, y)$, lo que da lugar a un total de 4 integraciones.
    3. Posteriormente, se aplica el submuestreo a las cuatro integraciones resultantes y se cuantifican utilizando 8 *bins* cada una. Esto da como resultado un total de $8 + 8 + 8 + 8 = 32$ *bins*.

    En resumen, el resultado de MIFH-Qm es un vector de características con la misma longitud que el MIFH original. Este enfoque se basa en la cuantificación individual de los mapas de bordes en lugar de la integración global, lo que permite capturar información detallada en cada etapa del proceso.

- MIFH-Qr sigue el mismo modelo de cuantificación utilizado en MIFH, que consiste en dividir los valores en 16 *bins*. Sin embargo, se aplica una regla de cuantificación específica para lograr una mejor representación de la información en los valores más pequeños de la integración, que se ha observado contienen detalles importantes. La regla de cuantificación se establece de la siguiente manera:

    - Los diez primeros valores se cuantifican en rangos de 0.05.
    - Los cuatro valores siguientes se cuantifican en rangos de 0.075.
    - Finalmente, los últimos dos valores se cuantifican en rangos de 0.1.

    MIFH-Qr busca evitar una generalización en los valores más pequeños de la integración, reconociendo que contienen información más detallada y relevante. La regla de cuantificación para obtener cada $M_1(x, y), M_2(x, y)$, y $M_3(x, y)$ a partir de cada $E_1(x, y), E_2(x, y)$, y $E_3(x, y)$ se muestra en Eq. (4.1).

    En general MIFH-Qr conserva la cuantificación en 16 bins utilizada en MIFH, pero adapta la regla de cuantificación para reflejar la importancia de los valores más

pequeños en la integración, mejorando así la representación de la información contenida en los mapas de bordes.

$$M(x,y) = \begin{cases} 0, & E(x,y) \in [0, 0.05] \\ 1, & E(x,y) \in [0.05, 0.10] \\ 2, & E(x,y) \in [0.10, 0.15] \\ 3, & E(x,y) \in [0.15, 0.20] \\ 4, & E(x,y) \in [0.20, 0.25] \\ 5, & E(x,y) \in [0.25, 0.30] \\ 6, & E(x,y) \in [0.30, 0.35] \\ 7 & E(x,y) \in [0.35, 0.40] \\ 8' & E(x,y) \in [0.40, 0.45] \\ 9, & E(x,y) \in [0.45, 0.50] \\ 10, & E(x,y) \in [0.50, 0.575] \\ 11, & E(x,y) \in [0.575, 0.650] \\ 12, & E(x,y) \in [0.650, 0.725] \\ 13, & E(x,y) \in [0.725, 0.80] \\ 14, & E(x,y) \in [0.8, 0.9] \\ 15, & E(x,y) \in [0.9, 1.0] \end{cases} \quad (4.1)$$

4.1.4 Multi-Integration Feature Histogram Weighted

Durante la detección de debilidades se encontró que el color de fondo de la imagen provocaba la recuperación de imágenes no deseadas. Ya que dentro de las imágenes recuperadas se encontraban resultados que únicamente contenían colores de fondo similares. Con el fin de minimizar el impacto del color de fondo en la recuperación, se planteó una modificación llamada (MIFH-W). Donde se atribuyen pesos a cada zona, con la finalidad de reducir la importancia de las características del fondo en la imagen.

MIFH-W se ha propuesto reconociendo que las regiones de interés en una imagen suelen contener información de bordes más significativa. Por lo tanto, se ha modificado la forma en que se construye el histograma de color en este descriptor. En lugar de tratar toda la imagen de manera uniforme, MIFH-W divide la imagen en nueve sub-imágenes y asigna pesos a cada una de ellas en función de la presencia de bordes. En otras palabras, si una región de la imagen contiene una mayor cantidad de información de bordes, el descriptor dará más peso a los colores presentes en esa región, aumentando gradualmente los pesos asignados a cada sub-imagen en incrementos de 0.022. En contraste, las regiones de fondo, que tienden a ser más homogéneas, recibirán un peso de 0.022, lo que equivale a un 2.2% de contribución al histograma de color. Por otro lado, las regiones consideradas de interés obtendrán un peso de 0.198, lo que representa un 19.8% de aportación al histograma de color.

MIFH-W garantiza que el vector de color esté compuesto principalmente por los colores presentes en las zonas consideradas de interés, reflejando así la importancia de estas

regiones en la representación del contenido de la imagen. Al mismo tiempo, se reduce la influencia de las regiones de fondo homogéneas en el descriptor.

4.2 Propuestas basadas en Microestructuras

El descriptor CMSD es propuesto como una mejora al descriptor MSD. El descriptor identifica microestructuras mediante el establecimiento de correlaciones entre la orientación de la textura, el color y las características de intensidad. Utiliza el espacio de color HSV y lo cuantifica en 72 *bins*, las características del borde cuantificadas en 6 *bins*, y las características de la luminosidad en 10 *bins*.

4.2.1 Oportunidades de mejora en Correlated MicroStructure Descriptor

A pesar de que el descriptor CMSD entrega buenos resultados en la recuperación de imágenes, se observó que contempla únicamente estructuras pequeñas, ya que utiliza una ventana 3×3, lo que podría afectar su desempeño en la recuperación. Asimismo, el descriptor CMSD únicamente considera la correlación de las microestructuras, y no los tipos de estructura encontrados en las imágenes. Por otro lado, el descriptor podría verse limitado al tipo de técnicas utilizadas para la obtención de los mapas de características.

4.2.2 Pyramid Correlated Microstructure Descriptor.

Se propuso una variante mejorar el rendimiento del descriptor CMSD mediante la integración de diversas escalas de estructuras utilizando un enfoque piramidal, como se ilustra en la Figura 4.1. El proceso piramidal consiste en generar múltiples versiones de la imagen original, reduciendo su escala a la mitad en cada nivel. Esto permite capturar estructuras de mayor tamaño en las imágenes escaladas. Esta versión mejorada se denomina "*Pyramid Correlated Microstructure Descriptor*" (PCMSD).

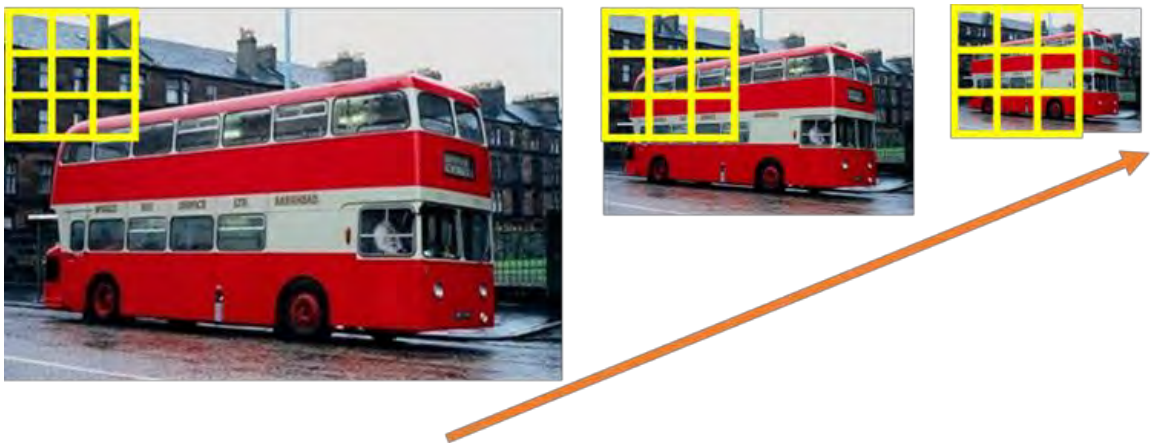


Figura 4.1 Submuestreo piramidal

El vector resultante de PCMSD se construye concatenando los vectores obtenidos al aplicar el descriptor CMSD a cada una de las imágenes de la pirámide generadas. Con una profundidad de tres niveles en la pirámide, se obtienen tres imágenes de diferentes tamaños

y la imagen original. Cada una de estas imágenes se describe utilizando el descriptor CMSD, generando un vector de longitud $88 \times 4 = 352$ valores en total.

En síntesis, la primera parte del vector PCMSD proporciona información sobre microestructuras más pequeñas, mientras que la última parte del vector ofrece información sobre microestructuras más grandes. Esto se debe a que las imágenes utilizadas para calcular estas partes del vector son significativamente más pequeñas, y la ventana de detección mantiene su tamaño original. El uso de diferentes submuestreos de la imagen permite capturar de manera más efectiva la información de las microestructuras en una variedad de escalas.

4.2.3 Correlated Microstructure Element's Descriptor

Considerando que, la obtención de las microestructuras requiere la detección de estructuras y que existen descriptores como SED [109], que utilizan la información de los elementos de estructuras para representar la imagen, se presentó una variante que utiliza los tipos de estructura y las microestructuras llamada "*Correlated Microstructure and Element's Descriptor*" (CMED). El descriptor se presenta como una mejora al descriptor CMSD [122], y SED [109], ya que mediante el uso de los elementos de estructuras propuestos y la detección en diferentes escalas, obtiene una representación superior en clases semánticas.

El descriptor se basa en dos teorías visuales. La teoría de integración de características presentada por Treisman [49], establece dos etapas de la visión "*Pre-attentive*" donde se detectan las características de bajo nivel y la etapa "*Attentive*", donde se recombinan las características para formar estructuras más complejas. La teoría de los "*Textons*" de Julesz [50], presenta el término *Textons* como las microestructuras universales por las que están compuestas las imágenes. De tal manera que el descriptor logra obtener la información de la relación entre la textura y el color, con el uso de la correlación de las microestructuras y los elementos de estructuras.

La diferencia entre los descriptores SED y CMSD, con respecto a CMED es el uso de ambas metodologías y el proceso para obtener el histograma del elemento. Dado que en lugar de utilizar tipos clásicos de estructuras basadas en la forma como SED, el descriptor propuesto utiliza una metodología novedosa basada en el número de elementos, que puede ser invariable a las transformaciones geométricas. La metodología de extracción de características puede ser descrita con los siguientes cinco pasos:

Paso 1. Se extraen las características de bajo nivel.

Paso 2. Los mapas de microestructura se generan utilizando las características obtenidas en la extracción.

Paso 3. Los elementos de la estructura se detectan utilizando los mapas de microestructura de cada característica.

Paso 4. Las correlaciones de las microestructuras se obtienen utilizando los tres mapas de características, lo que da como resultado tres mapas de correlación.

Paso 5. Los histogramas de los elementos de estructura detectados en cada característica se generan y concatenan para obtener un único histograma de estructuras. Por otra parte, se obtiene el histograma de cada mapa de correlación de microestructuras y se concatena en un solo histograma de correlación. Finalmente, ambos histogramas se utilizan para producir el vector descriptor CMED.

La Figura 4.2 muestra la arquitectura propuesta para el descriptor, el cual se basa en los descriptores CMSD, MSD y SED. Asimismo, los pasos que sigue el descriptor propuesto se detallan a continuación.

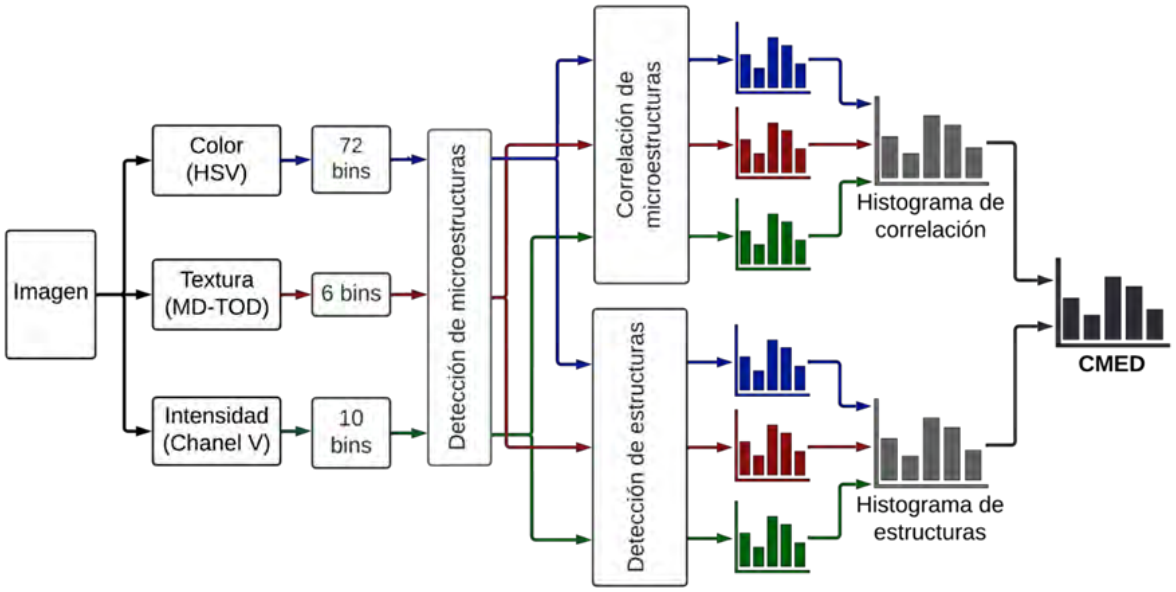


Figura 4.2 Proceso del descriptor propuesto CMED

4.2.3.1 Extracción de características de bajo nivel

El primer paso es extraer los mapas de características y cuantificarlos, para poder generalizar. La imagen se transforma del espacio de color RGB al espacio de color HSV, ya que el espacio de color HSV se reporta en la literatura como más similar a la percepción humana [122]. Para generar el mapa de color, la imagen en HSV, se cuantifica cada canal $H(x,y)$, $S(x,y)$ y $V(x,y)$, uniformemente con $Bh = 8$, $Bs = 3$ y $Bv = 3$, como se muestra en las Eq. (4.2) - (4.4). Utilizando los valores en cada canal cuantificado se obtiene el mapa de color $CM(x,y)$ el número de filas y columnas de la imagen original con $8 \times 3 \times 3 = 72$ valores distintos. Resultando así en una $matriz(x,y)$ con valores en el rango $[0 - 71]$, la obtención del mapa se muestra en Eq. (4.5).

$$Q_h(x,y) = H(x,y) \times (B_h / max_h) \qquad (4.2)$$

$$Q_s(x,y) = S(x,y) \times (B_s/max_s) \quad (4.3)$$

$$Q_v(x,y) = V(x,y) \times (B_v/max_v) \quad (4.4)$$

$$CM(x,y) = Q_h(x,y) \times (B_s \times B_v) + Q_s(x,y) \times B_v + Q_v(x,y) \quad (4.5)$$

El mapa de intensidad $IM(x,y)$ se obtiene cuantificando en *bins* de 10 el canal de intensidad V de la imagen en HSV. Considerando que el valor en el canal V está en el rango de cero a uno, el mapa de intensidad cuantificado se expresa como Eq. (4.6), donde B_I se establece en diez, y $V(x,y)$ como el valor en V, de las coordenadas x,y, en la imagen HSV. Dando como resultado una matriz con valores en el rango $[0 - 9]$.

$$IM(x,y) = V(x,y) \times B_I \quad (4.6)$$

Finalmente, el mapa de bordes se obtiene con la metodología propuesta en [122], que se denomina "*Multi-Dimensional Texture Orientation Detection*" (MD-TOD). La detección de bordes utiliza el Sobel en cuatro direcciones de bordes, $0°$, $45°$, $90°$ y $135°$, con las máscaras mostradas en la Figura 4.3. CMSD usa el espacio de color HSV, por lo que después de aplicar los filtros Sobel, el descriptor se transforma de coordenadas cilíndricas a cartesianas, como se muestra en Eq. (4.7) - (4.9), H_c, S_c y V_c son los nuevos valores del espacio HSV en coordenadas cartesianas.

0°

1	2	1
0	0	0
-1	-2	-1

45°

-2	-1	0
-1	0	1
0	1	2

90°

1	0	-1
2	0	-2
1	0	-1

135°

0	1	2
-1	0	1
-2	-1	0

Figura 4.3 Máscaras Sobel en 4 direcciones

$$H_c(x,y) = S(x,y) \times cos(H(x,y)) \quad (4.7)$$

$$S_c(x,y) = S(x,y) \times sin\big(H(x,y)\big) \quad (4.8)$$

$$V_c(x,y) = V(x,y) \tag{4.9}$$

Con las coordenadas cartesianas HSV, la metodología MD-TOD detecta los vectores de aristas diagonales denotados por $\widehat{d_{45°}}$ y $\widehat{d_{135°}}$, siendo la información de borde extraída usando los operadores de Sobel. El mapa de orientación del borde diagonal $Orimap_{diag}$, se obtiene mediante el ángulo entre dos vectores Eq. (4.10) - (4.14). $H_{45°}$, $S_{45°}$, y $V_{45°}$, son los bordes extraídos con la máscara Sobel $45°$ para cada canal. A su vez $H_{135°}$, $S_{135°}$, y $V_{135°}$, son los bordes detectados con el operador Sobel $135°$.

$$cos\left(\widehat{d_{45°}}, \widehat{d_{135°}}\right) = \frac{\widehat{d_{45°}} \times \widehat{d_{135°}}}{\left|\widehat{d_{45°}}\right|\left|\widehat{d_{135°}}\right|} \tag{4.10}$$

$$\widehat{d_{45°}} \times \widehat{d_{135°}} = H_{45°} \cdot H_{135°} + S_{45°} \cdot S_{135°} + V_{45°} \cdot V_{135°} \tag{4.11}$$

$$\left|\widehat{d_{45°}}\right| = \left(H_{45°}^2 + S_{45°}^2 + V_{45°}^2\right)^{\frac{1}{2}} \tag{4.12}$$

$$\left|\widehat{d_{135°}}\right| = \left(H_{135°}^2 + S_{135°}^2 + V_{135°}^2\right)^{\frac{1}{2}} \tag{4.13}$$

$$Orimap_{diag} = arccos\left(cos\left(\widehat{d_{45°}}, \widehat{d_{135°}}\right)\right) \tag{4.14}$$

La orientación de borde horizontal y vertical $Orimap_{hv}$, se extrae utilizando los vectores de bordes horizontal y vertical denotados por $\hat{h}$ y $\hat{v}$, Eq. (4.15) - (4.19), con H_h, S_h, V_h, siendo los bordes detectados para cada canal con Sobel Horizontal y H_v, S_v, V_v, los bordes detectados con Sobel Vertical.

$$cos\left(\hat{h}, \hat{v}\right) = \frac{\hat{h} \times \hat{v}}{\left|\hat{h}\right|\left|\hat{v}\right|} \tag{4.15}$$

$$\hat{h} \times \hat{v} = H_h \cdot H_v + S_h \cdot S_v + V_h \cdot V_v \tag{4.16}$$

$$\left|\widehat{d_h}\right| = \left(H_h^2 + S_h^2 + V_h^2\right)^{\frac{1}{2}} \tag{4.17}$$

$$\left|\widehat{d_v}\right| = \left(H_v^2 + S_v^2 + V_v^2\right)^{\frac{1}{2}} \tag{4.18}$$

$$Orimap_{hv} = arccos\left(cos(\hat{h}, \hat{v}\,)\right) \tag{4.19}$$

El mapa de orientación del borde $OM(x, y)$, se obtiene considerando las cuatro direcciones utilizando la Eq. (4.20). Donde B_O, es el nivel de cualificación establecido en 6. El resultado de la extracción de características de bajo nivel entrega tres mapas de características cuantificados, como se muestra en la Figura 4.4.

$$OM(x, y) = \frac{\left(Orimap_{hv} + Orimap_{diag}\right)}{2} \times \frac{B_O}{180} \tag{4.20}$$

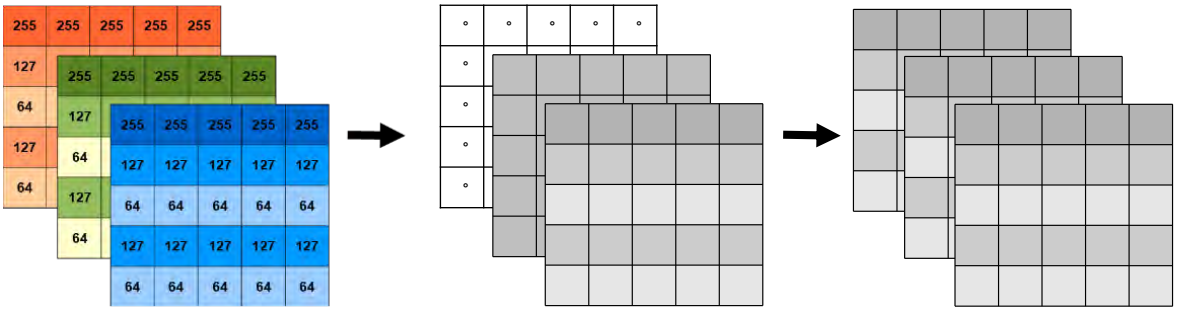


Figura 4.4 Paso 1 del descriptor propuesto CMED

4.2.3.2 Detección de microestructuras

Se sigue la metodología propuesta por [121] para llevar a cabo la detección de microestructuras. En la Figura 4.5 se ilustra un ejemplo del proceso utilizado para esta detección.

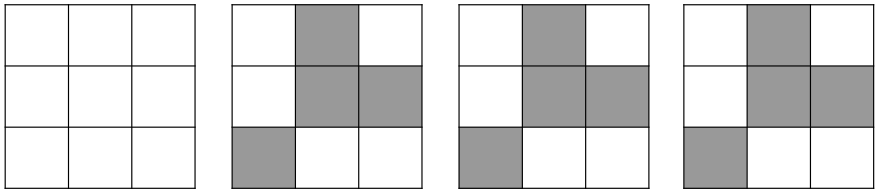


Figura 4.5 Detección de Microestructura fundamental

El método se aplica en cada mapa utilizando un bloque de tamaño 3×3. La detección se realiza a lo largo de todo el mapa, siguiendo una dirección de izquierda a derecha y de arriba a abajo, utilizando cuatro puntos de inicio diferentes $(0, 0), (0, 1), (1, 0)$ y $(1, 1)$, y un valor de desplazamiento de tres ($stride = 3$), como se muestra en la Figura 4.6.

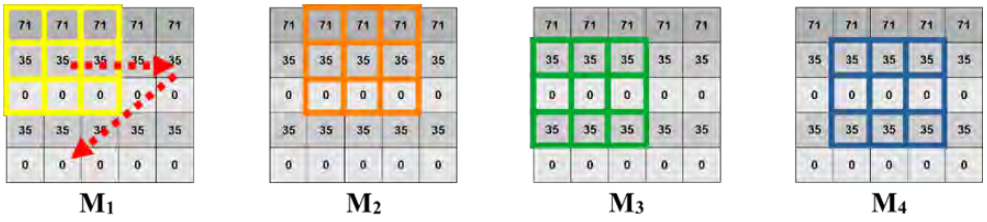


Figura 4.6 Mapas de Microestructuras

El proceso resulta en la obtención de cuatro mapas de microestructura distintos: M_1, M_2, M_3 y M_4, correspondientes a cada uno de los cuatro puntos de inicio mencionados previamente. Estos cuatro mapas de microestructura se combinan utilizando la Eq. (4.21), dando como resultado una imagen de microestructura, denotada como $M_T(x,y)$, ilustrada en la Figura 4.7. El procedimiento se repite para cada mapa de características, lo que conduce a la generación de tres imágenes de microestructuras, una para cada mapa de características, como se muestra en la Figura 4.8.

$$M_T(x,y) = Max\big(M_1(x,y), M_2(x,y), M_3(x,y), M_4(x,y)\big) \qquad (4.21)$$

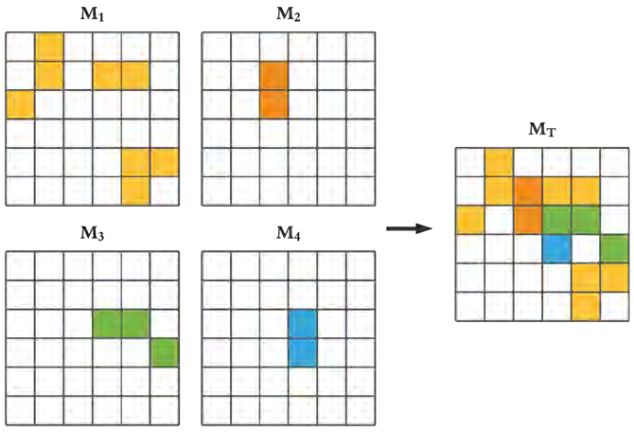


Figura 4.7 Imagen de microestructura

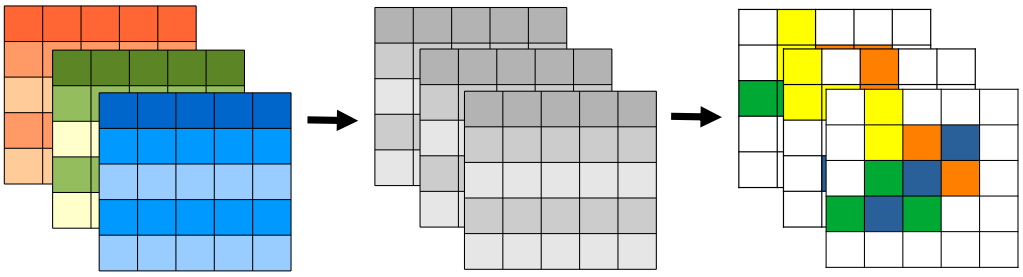


Figura 4.8 Paso 2 del descriptor propuesto CMED

4.2.3.3 Detección de estructuras

Partiendo de la idea presentada en [109], se detectan los tipos de estructuras presentes en las microestructuras. Sin embargo, a diferencia de los tipos utilizados en la literatura, que se basan en la dirección o forma de la estructura como se muestra en la Figura 4.9, para el descriptor se propone una nueva forma de categorizar los tipos.

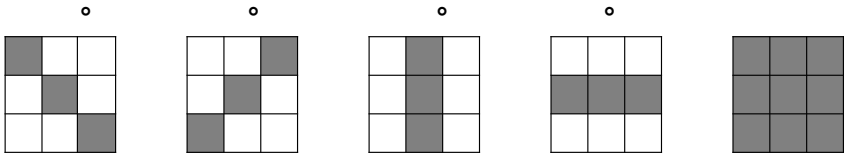


Figura 4.9 Ejemplos de tipos de estructuras clásicas

La nueva categorización se basa en la cantidad de elementos presentes en la estructura, como se muestra en la Figura 4.10. Este enfoque proporciona tolerancia a la rotación y un número reducido de tipos. En este caso, se utilizan ocho tipos, o nueve si se consideran las estructuras sin elementos, a diferencia de las estructuras clásicas que pueden tener incluso más de 16 tipos. De esta manera, las microestructuras previamente detectadas se etiquetan según la cantidad de elementos que poseen en la estructura correspondiente.

Para obtener los elementos de las estructuras, la detección comienza en el mismo origen que cada mapa de microestructuras, con un bloque de 3×3 que se desplaza de izquierda a derecha y de arriba a abajo con un $stride = 3$. El tipo de estructura se asigna según el número de elementos y se guarda en un vector con longitud ns, donde ns es el número de tipos de estructura. Finalmente, se obtienen cuatro histogramas, uno por cada mapa de microestructuras, los cuatro histogramas se suman y dividen entre cuatro para obtener un único vector, por cada una de las características.

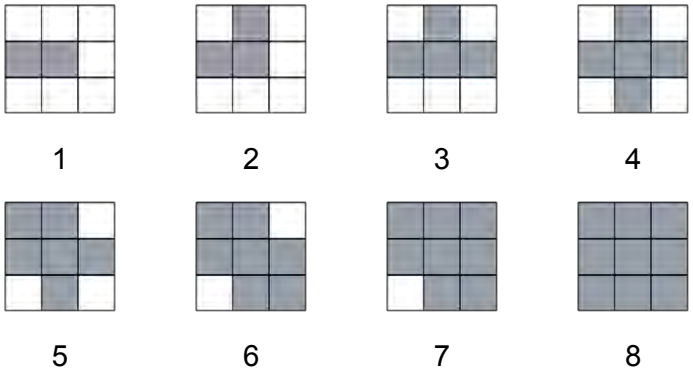


Figura 4.10 Tipos de estructuras basadas en cantidad de elementos

No obstante, la representación puede resultar demasiado general. Como se ilustra en la Figura 4.12, existen numerosas formas diferentes que comparten el mismo número de elementos. Por esta razón, se llevó a cabo una búsqueda exhaustiva para identificar diversas estructuras posibles, teniendo en cuenta tanto la rotación como la reflexión.

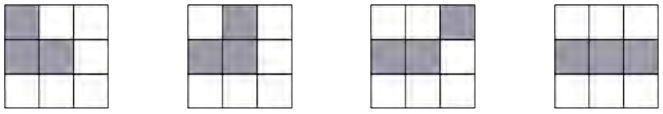


Figura 4.11 Diferentes formas de estructuras con dos elementos

En particular, se agruparon las estructuras que surgían de rotaciones en la misma categoría, como se visualiza en la Figura 4.12. De manera análoga, las estructuras derivadas de reflexión o transformaciones tipo espejo se trataron como una única categoría, como se muestra en la Figura 4.13. Esto se hizo para preservar la tolerancia a tales transformaciones.

0° 90° 180° 270°

Figura 4.12 Estructura de cuatro elementos en diferentes orientaciones

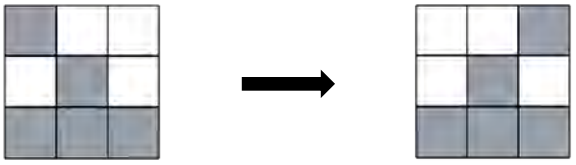


Figura 4.13 Reflexión o transformación tipo espejo

Para asignar un tipo específico a cada estructura, se siguió un proceso detallado. En primer lugar, se extrajeron los ocho vecinos del bloque 3×3 y se dispusieron en un vector. Este vector se creó asignando "1 a las posiciones con un elemento y "0" a las posiciones sin elemento. Luego, siguiendo el enfoque de LBP-U [152], se contaron las transiciones, es decir, los cambios de "1" a "0" y de "0" a "1", lo que se representa como C_t. Además, se determina la distancia máxima entre dos elementos C_d y la longitud máxima de elementos continuos C_l. También se toma en cuenta el número de elementos simétricos diferentes C_s siguiendo la Eq. (4.22), donde V es el vector generado a partir de los ocho vecinos de la estructura. Es importante destacar que, para obtener todos los coeficientes, se considera que la posición "ocho" en el vector precede a la posición "uno". En última instancia, estos cuatro valores proporcionaron un código de identificación único para cada tipo de estructura.

$$C_s = \sum_{i=1}^{8} |V(i) - V(i+4)| \qquad (4.22)$$

A modo de ilustración, si consideramos el vector de una estructura $V_e = [1,1,1,0,1,0,0,0]$, se pueden calcular los siguientes coeficientes: $C_t = 4, C_d = 3, C_l = 3$ y $C_s = 2$. De esta manera, el código de identificación de la estructura sería 4332, lo que denota que todas las estructuras detectadas con ese código se clasificarían del mismo tipo, el cual se estableció como tipo 12. En total, teniendo en cuenta estructuras con ocho y cero elementos, se identificaron un conjunto de treinta tipos diferentes las cuales se denominaron estructuras fundamentales, como se muestra en la Figura 4.14.

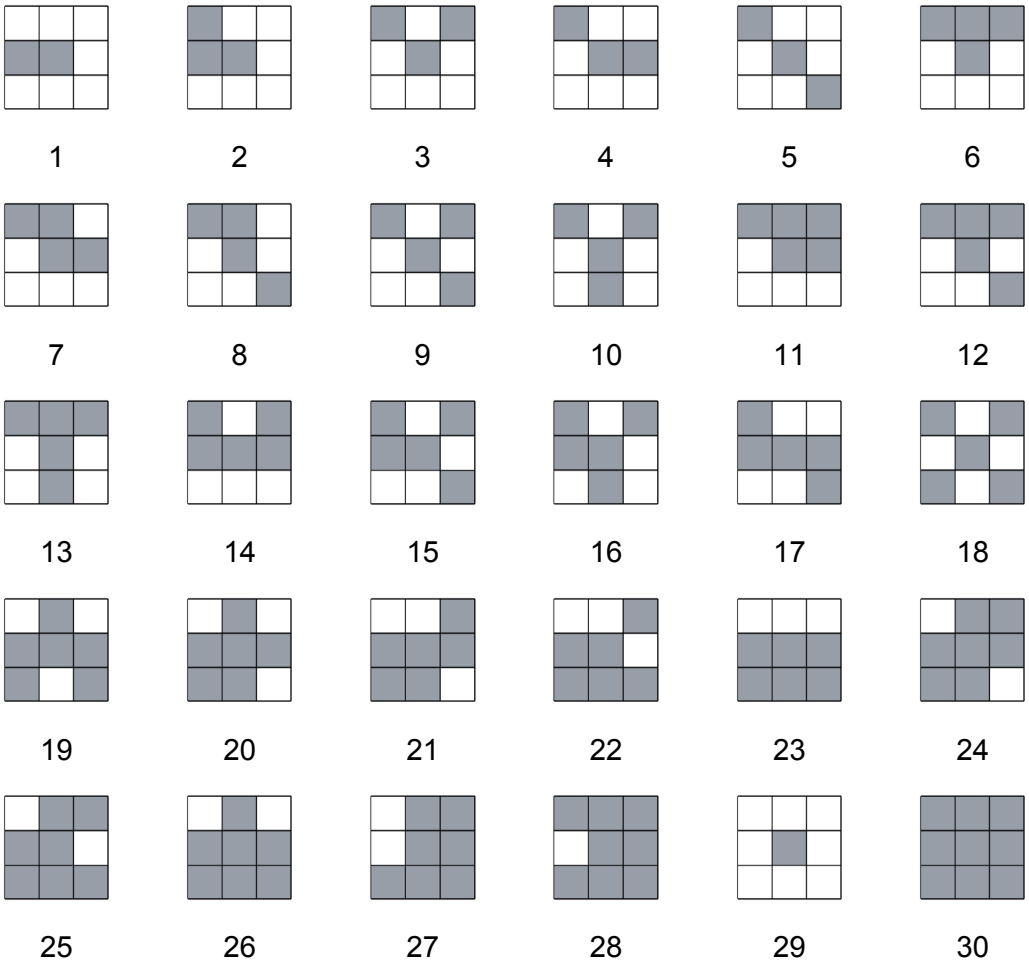


Figura 4.14 Estructuras fundamentales

4.2.3.4 Correlación de microestructuras

Para obtener una descripción que contemple la relación entre características de bajo nivel, el descriptor propuesto sigue el proceso de correlación de microestructuras presentado en [122], donde se obtienen para cada característica un Mapa de correlación de Micro-Característica, utilizando dos imágenes de microestructuras para extraer información de

uno de los mapas de características. Las imágenes de microestructura se combinan para obtener un mapa de microestructura correlacionado $M_c(x,y)$. Considerando el método propuesto en [122] $M_c(x,y)$ se puede obtener como se muestra en la Eq. (4.23) el cual se nombró para este trabajo correlación "OR", ya que se obtendrá la microestructura si está presente en el primero o segundo mapa. Así mismo, se propone una obtención de la correlación que se nombró "AND", que se obtiene mediante Eq. (4.24), donde solo se consideran las estructuras que se presentan en ambos mapas.

$$M_C(x,y) = \max\left(M_T^1(x,y), M_T^2(x,y)\right) \tag{4.23}$$

$$M_C(x,y) = M_T^1(x,y) \times M_T^2(x,y) \tag{4.24}$$

Finalmente, se utiliza el $M_c(x,y)$ para extraer información del mapa de características, tomando los valores ubicados en las microestructuras. El resultado da un mapa de Micro-características. Todo el proceso se repite para cada característica, lo que da como resultado el mapa de Micro-color, el mapa de Micro-orientación y el mapa de Micro-intensidad. Un ejemplo ilustrativo del proceso para obtener el mapa de Micro-Orientación utilizando una correlación "OR" se presenta en la Figura 4.15, utilizando las imágenes de microestructura $M_T(x,y)$, de intensidad y color para extraer información del mapa de orientación.

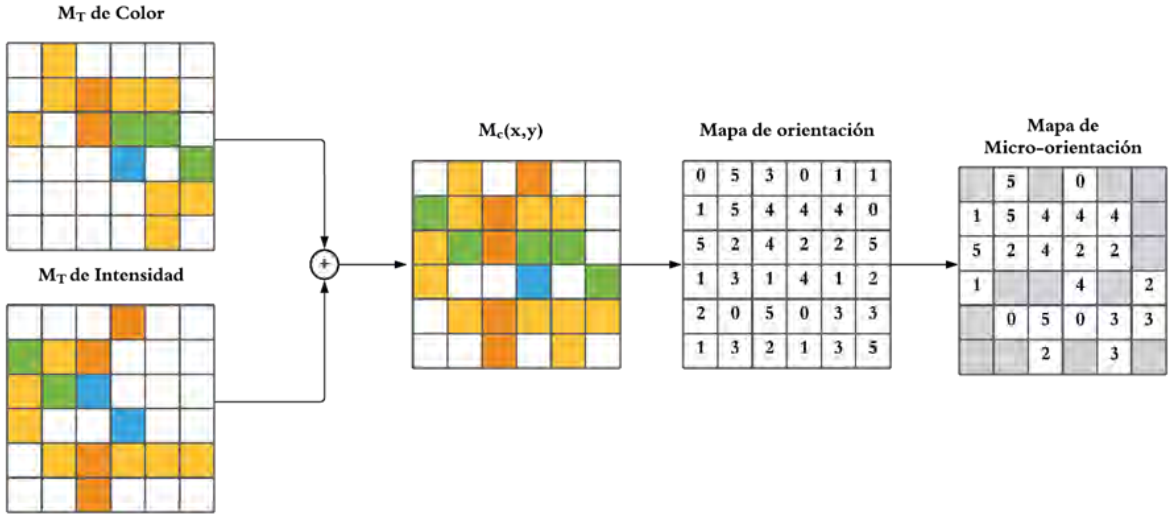


Figura 4.15 Construcción del mapa de Micro-orientación

4.2.3.5 Representación de características

El descriptor es representado mediante un histograma. En primera instancia se concatenan los histogramas de estructuras. En el caso de considerar la estructura con cero elementos para generar el histograma de los tipos de estructuras, es decir $ns = 9$, se obtienen un total de $9 \times 3 = 27$ valores. Asimismo, se considera el histograma de correlación de microestructuras de cada característica, obteniendo así un vector con valores $72 + 6 +$

$10 = 88$. Finalmente, el vector del descriptor tiene una longitud de $27 + 88 = 115$. De esta manera el descriptor CMED contiene información sobre la ocurrencia de cada tipo de estructuras y microestructuras. El vector CMED se define en Eq. (4.25). Donde H_s^C, H_s^O y H_s^I, son los histogramas de los elementos de la estructura para color, orientación e intensidad, respectivamente. H_m^C, H_m^O y H_m^I los histogramas del mapa Micro-color, Micro-orientación y Micro-intensidad, respectivamente.

$$CMED = [H_s^C, H_s^O, H_s^I, H_m^C, H_m^O, H_m^I] \tag{4.25}$$

4.2.4 Fractional Correlated Microestructure Element's Descriptor

Considerando los resultados encontrados en la literatura sobre el cálculo fraccionario o cálculo de orden arbitrario, y a fin de mejorar el descriptor propuesto en términos de representación y tolerancia a la presencia de ruido en las imágenes. Se propusieron dos variantes utilizando la detección de bordes basados en cálculo fraccionario para obtener la descripción de la textura mediante la orientación del borde. Se implementaron dos metodologías para la detección de Sobel fraccionario, utilizando la definición de Grunwald Letniko (GL) [153] y la definición de Caputo–Fabrizio (CF) [154]. Las dos metodologías se adaptaron para trabajar en imágenes en el espacio de color HSV siguiendo la técnica de MD-TOD y se utilizó con un $\alpha = 1.5$, es decir se obtuvo la una y media derivada.

La primera variante llamada "*Correlated Microestructure Elements Descriptor with Grunwald Letniko*" (CMED-GL), toma como base el método propuesto en [153] para detectar los bordes utilizando la Eq. (2.4). Los autores parten de la definición de GL y proponen dos máscaras una en x y otra en y de un tamaño de 5×5, basadas en Sobel como se muestra en la Figura 4.16.

Máscara x

$\frac{\alpha^2-\alpha}{2}$	$-\alpha$	0	α	$\frac{\alpha-\alpha^2}{2}$
$2\frac{\alpha^2-\alpha}{2}$	-2α	0	2α	$2\frac{\alpha-\alpha^2}{2}$
$3\frac{\alpha^2-\alpha}{2}$	-3α	0	3α	$3\frac{\alpha-\alpha^2}{2}$
$2\frac{\alpha^2-\alpha}{2}$	-2α	0	2α	$2\frac{\alpha-\alpha^2}{2}$
$\frac{\alpha^2-\alpha}{2}$	$-\alpha$	0	α	$\frac{\alpha-\alpha^2}{2}$

Máscara y

$\frac{\alpha^2-\alpha}{2}$	$2\frac{\alpha^2-\alpha}{2}$	$3\frac{\alpha^2-\alpha}{2}$	$2\frac{\alpha^2-\alpha}{2}$	$\frac{\alpha^2-\alpha}{2}$
$-\alpha$	-2α	-3α	-2α	$-\alpha$
0	0	0	0	0
α	2α	3α	2α	α
$\frac{\alpha-\alpha^2}{2}$	$2\frac{\alpha-\alpha^2}{2}$	$3\frac{\alpha-\alpha^2}{2}$	$2\frac{\alpha-\alpha^2}{2}$	$\frac{\alpha-\alpha^2}{2}$

Figura 4.16 Máscara fraccional GL horizontal y vertical [153]

Sin embargo, el método MD-TOD requiere de las máscaras para las aristas diagonales las cuales se obtuvieron basándose en las máscaras x y y, quedando como se muestra en la Figura 4.17.

Máscara $45°$

$3\frac{\alpha^2-\alpha}{2}$	$2\frac{\alpha^2-\alpha}{2}$	$\frac{\alpha^2-\alpha}{2}$	$-\alpha$	0
$2\frac{\alpha^2-\alpha}{2}$	-3α	-2α	0	α
$\frac{\alpha^2-\alpha}{2}$	-2α	0	2α	$\frac{\alpha-\alpha^2}{2}$
$-\alpha$	0	2α	3α	$2\frac{\alpha-\alpha^2}{2}$
0	α	$\frac{\alpha-\alpha^2}{2}$	$2\frac{\alpha-\alpha^2}{2}$	$3\frac{\alpha-\alpha^2}{2}$

Máscara $135°$

0	$-\alpha$	$\frac{\alpha^2-\alpha}{2}$	$2\frac{\alpha^2-\alpha}{2}$	$3\frac{\alpha^2-\alpha}{2}$
α	0	-2α	-3α	$2\frac{\alpha^2-\alpha}{2}$
$\frac{\alpha-\alpha^2}{2}$	2α	0	-2α	$\frac{\alpha^2-\alpha}{2}$
$2\frac{\alpha-\alpha^2}{2}$	3α	2α	0	$-\alpha$
$3\frac{\alpha-\alpha^2}{2}$	$2\frac{\alpha-\alpha^2}{2}$	$\frac{\alpha-\alpha^2}{2}$	α	0

Figura 4.17 Máscaras fraccionales GL diagonales [153]

Las cuatro máscaras fraccionales se utilizan en sustitución a las máscaras tradicionales Sobel, llamado Sobel-GL. Por lo que se extrae la información del borde de la imagen HSV en coordenadas cartesianas utilizando MD-TOD utilizando Sobel-GL. Es decir, el mapa de orientación del borde diagonal $Orimap_{diag}$, se obtiene mediante Eq. (4.10) - (4.14). Donde, $H_{45°}$, $S_{45°}$, y $V_{45°}$, son los bordes extraídos con la máscara fraccional GL de $45°$ para cada canal, y $H_{135°}$, $S_{135°}$, y $V_{135°}$, los bordes detectados con la máscara GL de $135°$. Asimismo, el mapa de orientación $Orimap_{hv}$, se extrae utilizando la Eq. (4.15) - (4.19), con H_h, S_h, V_h, siendo los bordes detectados para cada canal con la máscara fraccional GL x y H_v, S_v, V_v, los bordes detectados con la máscara y.

En cuanto a la variante basada en Caputo–Fabrizio, llamada "*Correlated Microestructure Elements Descriptor with Caputo–Fabrizio*" (CMED-CF). La extracción del borde se implementó utilizando lo que llamamos Sobel-CF como se propone en [155]. Donde a partir de la Eq. (2.3), obtienen la Eq. (4.26), utilizada para la extracción de borde en x, y la Eq. (4.27) para y, siendo $f_1(t,x) = f(x+1) - f(x-1)$, $f_2(t,y) = f(y+1) - f(y-1)$, $M(1) = M(0) = 1$ y α el orden de la derivada.

$$h_{x_{n+1}} = h_{x_n} + \left(\frac{1-\alpha}{M(\alpha)} + \frac{3\alpha h}{2M(\alpha)}\right) f_1(t_n, x_n) + \left(\frac{1-\alpha}{M(\alpha)} + \frac{\alpha h}{2M(\alpha)}\right) f_1(t_{n-1}, x_{n-1}) \qquad (4.26)$$

$$h_{y_{n+1}} = h_{y_n} + \left(\frac{1-\alpha}{M(\alpha)} + \frac{3\alpha h}{2M(\alpha)}\right) f_2(t_n, y_n) + \left(\frac{1-\alpha}{M(\alpha)} + \frac{\alpha h}{2M(\alpha)}\right) f_2(t_{n-1}, y_{n-1}) \qquad (4.27)$$

Las cuatro máscaras obtenidas basadas en Eq. (4.26) – (4.27), se muestra en la Figura 4.18. En este caso, las máscaras obtenidas con CF (máscara fraccional) se utilizan en lugar de las máscaras Sobel tradicionales del método MD-TOD. Es decir, $H_{45°}$, $S_{45°}$, y $V_{45°}$, representan los bordes extraídos con la máscara fraccional CF de $45°$. Mientras tanto, $H_{135°}$, $S_{135°}$, y $V_{135°}$, corresponden a los bordes detectados con la máscara CF de $135°$. Además, H_h, S_h, V_h, representan los bordes detectados para cada canal utilizando la máscara fraccional CF en la dirección horizontal, mientras que H_v, S_v, V_v, representan los bordes obtenidos con la máscara CF en la dirección vertical.

Máscara x

$\frac{\alpha}{2}-1$	0	$1-\frac{\alpha}{2}$
$\frac{-\alpha}{2}-1$	0	$1+\frac{\alpha}{2}$
$\frac{\alpha}{2}-1$	0	$1-\frac{\alpha}{2}$

Máscara y

$\frac{\alpha}{2}-1$	$\frac{\alpha}{2}-1$	$\frac{\alpha}{2}-1$
0	0	0
$1-\frac{\alpha}{2}$	$1+\frac{\alpha}{2}$	$1-\frac{\alpha}{2}$

Máscara $45°$

$\frac{-\alpha}{2}-1$	$\frac{\alpha}{2}-1$	0
$\frac{\alpha}{2}-1$	0	$1-\frac{\alpha}{2}$
0	$1-\frac{\alpha}{2}$	$1+\frac{\alpha}{2}$

Máscara $135°$

0	$\frac{\alpha}{2}-1$	$\frac{-\alpha}{2}-1$
$1-\frac{\alpha}{2}$	0	$\frac{\alpha}{2}-1$
$1+\frac{\alpha}{2}$	$1-\frac{\alpha}{2}$	0

Figura 4.18 Máscaras obtenidas mediante la definición de Caputo–Fabrizio [155]

5 Validación y experimentación

En esta sección se presentan los experimentos realizados, para evaluar el desempeño de las variantes propuestas en la recuperación de imágenes por contenido. Comparando los resultados entre las diferentes variantes propuestas y las técnicas en el estado del arte, al igual que con descriptores del estándar MPEG-7.

5.1 Construcción del entorno

Los experimentos se realizaron en Windows 10, con el software MATLAB en su versión R2020a, asimismo se ocupó el entorno de *Jupiter Notebooks* para la implementación de modelos previamente entrenados de CNN. Donde se utilizó la librería *Pythorch* y el módulo *Torchvision.* Se utilizaron los conjuntos de imágenes: Corel-1k; Corel-5k; y Caltech-101. Del conjunto de imágenes Caltech-101, se tomaron solo 100 de las 101 clases ya que no se consideraron las imágenes en escala de grises. Asimismo, como alternativa al conjunto de imágenes Corel-10k, el cual, es utilizado en gran parte de los trabajos revisados, se agregó el conjunto de imágenes "*Corel Database for Content based Image Retrieval*" [156], llamado en este trabajo (Corel-CBIR). Se tomaron 10 imágenes aleatorias por clase como imagen de consulta, utilizando la función *rand* de Matlab dando un total de 100 consultas por descriptor con Corel-1k; 500 con Corel-5k; y 1,000 con Caltech-101; y 800 con Corel-CBIR. Por fines prácticos y considerando que las personas buscan tener un resultado en las primeras imágenes recuperadas se establecieron las primeras evaluaciones en $K = 12$, es decir las primeras 12 imágenes más relevantes de cada consulta. Sin embargo, para obtener una evaluación completa, las ultimas comparaciones se realizaron con variaciones de K, en el rango $[1 - 100]$. Asimismo, para las evaluaciones realizadas con la métrica ANMRR se utilizaron todas las imágenes recuperadas con la finalidad de utilizar la métrica de manera correcta.

5.2 Experimentos

Los experimentos comienzan con la evaluación de las variantes, comparándolas entre las diferentes propuestas, en contraste con el descriptor o los descriptores tomados como base. Los experimentos continúan con una evaluación de la variante con mejor desempeño en contraste a los descriptores del estándar MPEG-7, los descriptores del estado del arte, Asimismo, se presenta una evaluación en contraste a descriptores

basado en características profundas. Los experimentos finalizan con el análisis de los resultados obtenidos.

5.2.1 Evaluación del rendimiento de las variantes

La primera parte de la evaluación detalla los resultados obtenidos con las variantes del descriptor MIFH, seguido de las variantes y resultados obtenidos con el descriptor CMSD.

5.2.1.1 Evaluación de las variantes del descriptor MIFH

Con la finalidad de subsanar más de una debilidad, se decidió combinar las variantes con posibilidad de hacerlo, asimismo se consideraron las variantes con mejor desempeño. El resultado son cuatro variantes, una de ellas es llamada MIFH-RQ, que cuantifica los mapas de borde como lo hace MIFH-Qm y reduce el vector eliminando los últimos valores como lo hace MIFH-R1. Otra combinación es la variante MIFH-RW, que asigna pesos a cada zona de la imagen como MIFH-W y reduce el vector final eliminando los últimos valores como R-MIFH-1. Asimismo, la variante MIFH-QW, asigna pesos a cada zona para la descripción como MIFH-W y cambia la cuantificación de los mapas siguiendo el proceso de MIFH-Qm. Finalmente, la variante MIFH-RQW, combina los procesos de las tres variantes MIFH-R1, MIFH-Qm y MIFH-W.

Los resultados obtenidos con las variantes propuestas para el descriptor MIFH se muestran de la Tabla 5.1 a la Tabla 5.4. Los resultados obtenidos con el descriptor original son resaltados en amarillo y los resultados de las variantes que superan o mantienen la evaluación con respecto al descriptor MIFH original son resaltados en color verde. Se observó que las variantes mejoran en algunas métricas y en ciertos conjuntos, sin embargo, no es posible obtener una variante estable que supere al descriptor original, lo que posiblemente se deba a que algunas de las variantes, a pesar de reforzar una debilidad. Un ejemplo se muestra en la Figura 5.1, donde se presenta la recuperación con MIFH y en la Figura 5.2 de una recuperación con MIFH-W. Se puede observar que en una clase donde el descriptor MIFH se desempeña muy bien ya que la mayoría de las imágenes tienen el mismo fondo blanco, afecta a la variante MIFH-W, ya que, al no considerar relevante el fondo, recupera en mejor posición las imágenes no deseadas, que tienen un color similar al objeto de interés.

Tabla 5.1 Evaluación de variantes MIFH con Corel-1k

	P	R	MAP	ANMRR
MIFH	75,25%	9,03%	8,32%	1,44%
MIFH-R1	75,67%	9,08%	8,35%	1,41%
MIFH-R2	75,42%	9,05%	8,34%	1,43%
MIFH-Qm	75,33%	9,04%	8,21%	1,44%
MIFH-Qr	74,17%	8,90%	8,21%	1,50%
MIFH-W	74,58%	8,95%	8,25%	1,47%
MIFH-RQ	74,92%	8,99%	8,19%	1,46%
MIFH-RW	74,92%	8,99%	8,29%	1,46%
MIFH-QW	74,42%	8,93%	8,12%	1,49%
MIFH-RQW	74,92%	8,99%	8,21%	1,46%

Tabla 5.2 Evaluación de variantes MIFH con Corel-5k

	P	R	MAP	ANMRR
MIFH	28,78%	3,45%	2,43%	4,22%
MIFH-R1	28,92%	3,47%	2,44%	4,22%
MIFH-R2	28,82%	3,46%	2,44%	4,22%
MIFH-Qm	29,27%	3,51%	2,46%	4,20%
MIFH-Qr	28,77%	3,45%	2,43%	4,22%
MIFH-W	29,27%	3,51%	2,48%	4,19%
MIFH-RQ	29,47%	3,54%	2,48%	4,18%
MIFH-RW	29,17%	3,50%	2,48%	4,20%
MIFH-QW	29,42%	3,53%	2,48%	4,19%
MIFH-RQW	29,68%	3,56%	2,50%	4,17%

Tabla 5.3 Evaluación de variantes MIFH con Corel-CBIR

	P	R	MAP	ANMRR
MIFH	32,04%	3,45%	2,66%	1,57%
MIFH-R1	31,96%	3,44%	2,67%	1,57%
MIFH-R2	31,97%	3,44%	2,65%	1,57%
MIFH-Qm	31,56%	3,39%	2,60%	1,58%
MIFH-Qr	31,32%	3,38%	2,60%	1,58%
MIFH-W	31,92%	3,42%	2,65%	1,57%
MIFH-RQ	31,58%	3,39%	2,59%	1,58%
MIFH-RW	31,94%	3,43%	2,65%	1,57%
MIFH-QW	31,02%	3,32%	2,54%	1,60%
MIFH-RQW	31,18%	3,34%	2,55%	1,59%

Tabla 5.4 Evaluación de variantes MIFH con Caltech-101

	P	R	MAP	ANMRR
MIFH	10,51%	1,58%	1,03%	4,57%
MIFH-R1	10,46%	1,57%	1,03%	4,57%
MIFH-R2	10,42%	1,56%	1,03%	4,57%
MIFH-Qm	10,60%	1,61%	1,05%	4,56%
MIFH-Qr	10,46%	1,59%	1,03%	4,57%
MIFH-W	10,46%	1,58%	1,04%	4,57%
MIFH-RQ	10,52%	1,59%	1,03%	4,57%
MIFH-RW	10,51%	1,58%	1,04%	4,57%
MIFH-QW	10,56%	1,62%	1,07%	4,56%
MIFH-RQW	10,56%	1,61%	1,05%	4,56%

Los resultados mostraron que MIFH-Qm fue la variante más estable en promedio, ya que obtiene mejores resultados que el MIFH original en tres de los cuatro conjuntos de datos de imágenes, sin embargo, sería necesario experimentar con diferentes *bins* de cuantificación para mejorar los resultados incluso en el conjunto de imágenes Corel-CBIR. De este modo, podría considerarse una posible opción para mejorar el rendimiento del descriptor.

Figura 5.1 Recuperación de dinosaurio con MIFH

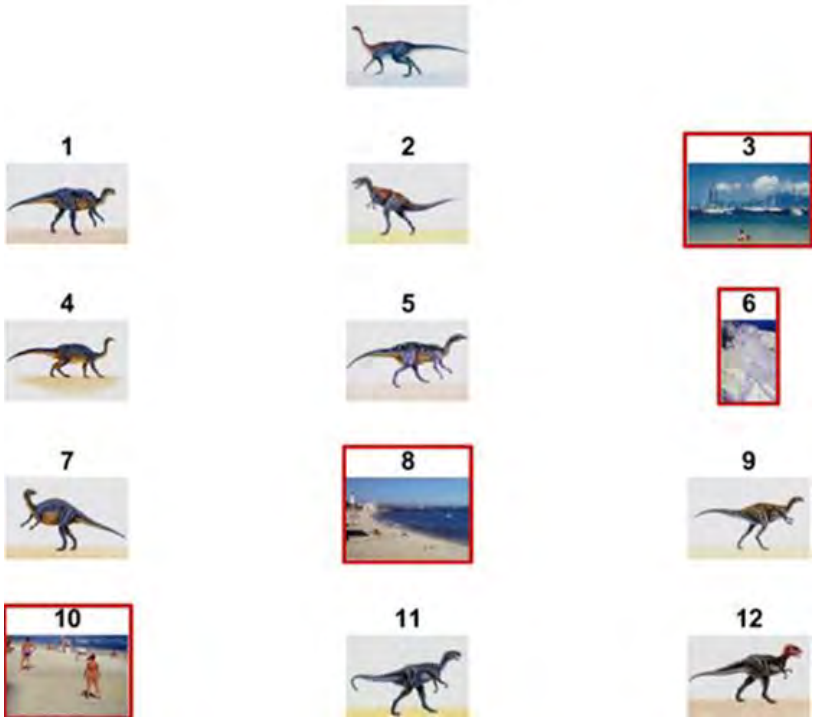


Figura 5.2 Recuperación de dinosaurio con MIFH-W

5.2.1.2 Evaluación de las variantes del descriptor CMSD

Considerando que la medida de similitud utilizada puede afectar el rendimiento del descriptor, y que el descriptor CMSD da gran importancia al color, se decidió implementar un ajuste de la medida de similitud *Manhattan* Eq. (2.22). La modificación propuesta se estableció como D_3, y se puede expresar como se muestra en Eq.(5.1). D_3

calcula la suma de las distancias para cada vector de características, dividiéndolas por el tamaño de la longitud del vector. Donde T se establece como el vector de una imagen del conjunto de imágenes a recuperar y Q, el vector de la imagen de consulta. Siendo N_f la longitud del vector de características, T^c y Q^c el vector de cada una de sus C características, y N_f^c su longitud. En el caso de CMSD se establece como $C = 3$, ya que considera borde, color e intensidad.

$$L_{c3}(T,Q) = \sum_{c=1}^{C} \frac{\sum_{i=1}^{N_f^c} |T_i^c - Q_i^c|}{N_f^c} \qquad (5.1)$$

Asimismo, se consideraron tres formas diferentes de obtener el vector de características durante el submuestreo piramidal. En la primera se obtiene el vector promedio de todos los vectores obtenidos a lo largo del submuestreo, por lo que se obtiene un vector de longitud 88. La segunda forma propuesta para la obtención del vector es concatenando cada uno de los vectores obtenidos en el proceso piramidal es decir si la profundidad del proceso piramidal es uno, el resultado será el mismo que el descriptor original. De esta manera el vector resultante es igual a $N_D \times 88$, donde N_D es la profundidad del submuestreo piramidal. Finalmente, considerando que se requiere menos detalles de color en cada submuestreo piramidal, ya que se trata de estructuras mayores, se modificaron los *bins* de cuantificación del mapa de color, con la finalidad de reducir el vector de características. Es decir, los *bins* de cuantificación del canal H decrecen en uno en cada submuestreo. Por lo que, la profundidad máxima de la imagen en el submuestreo piramidal es ocho.

Considerando que la profundidad máxima para uno de los métodos es ocho se estableció una profundidad de cuatro para comparar su rendimiento en la recuperación. Por lo que la primera variante obtiene un vector de longitud 88, la segunda variante una longitud de $4 \times 88 = 352$ y la tercera variante obtiene un vector de $88 + 79 + 70 + 61 = 298$ valores en el vector resultante. Los resultados se muestran de la Tabla 5.5 a la Tabla 5.8.

Tabla 5.5 Evaluación de variante CMSD con Corel-1k

	Longitud del vector	Medida de similitud	P	R	MAP	ANMRR
CMSD	88	Manhattan	77,50%	9,30%	8,66%	1,31%
PCMSD	88		77,08%	9,25%	8,68%	1,33%
	352		76,08%	9,13%	8,49%	1,39%
	298		79,08%	9,49%	8,91%	1,21%
	88	Propuesta	76,83%	9,22%	8,47%	1,35%
	352		75,25%	9,03%	8,25%	1,44%
	298		76,42%	9,17%	8,52%	1,37%
CMED	88	Manhattan	78,58%	9,43%	8,78%	1,24%
		Propuesta	78,25%	9,39%	8,70%	1,26%

Tabla 5.6 Evaluación de variante CMSD con Corel-5k

	Longitud del vector	Medida de similitud	P	R	MAP	ANMRR
CMSD	88	Manhattan	32,32%	3,88%	2,84%	4,01%
PCMSD	88		30,95%	3,71%	2,72%	4,09%
	352		29,97%	3,60%	2,60%	4,15%
	298		32,22%	3,87%	**2,89%**	4,02%
	88	Propuesta	25,37%	3,04%	2,12%	4,43%
	352		23,17%	2,78%	1,83%	4,56%
	298		26,97%	3,24%	2,28%	4,33%
CMED	88	Manhattan	32,75%	3,93%	2,91%	3,98%
		Propuesta	31,87%	3,82%	2,80%	4,04%

Tabla 5.7 Evaluación de variantes CMSD con Corel-CBIR

	Longitud del vector	Medida de similitud	P	R	MAP	ANMRR
CMSD	88	Manhattan	37,27%	3,99%	3,27%	1,45%
PCMSD	88		37,04%	3,99%	3,23%	1,45%
	352		34,67%	3,73%	2,96%	1,50%
	298		37,34%	4,02%	3,27%	**1**,44%
	88	Propuesta	37,27%	3,99%	3,27%	1,45%
	352		37,04%	3,99%	3,23%	1,45%
	298		34,67%	3,73%	2,96%	1,50%
CMED	88	Manhattan	39,75%	4,24%	3,51%	1,39%
		Propuesta	40,56%	4,32%	3,59%	1,37%

Tabla 5.8 Evaluación de variantes CMSD con Caltech-101

	Longitud del vector	Medida de similitud	P	R	MAP	ANMRR
CMSD	88	Manhattan	10,13%	1,59%	1,06%	4,57%
PCMSD	88		11,36%	1,85%	1,21%	4,51%
	352		10,90%	1,76%	1,15%	4,53%
	298		11,64%	1,87%	1,23%	4,50%
	88	Propuesta	13,57%	2,25%	1,55%	4,41%
	352		12,44%	2,00%	1,37%	4,47%
	298		13,64%	2,25%	1,55%	4,41%
CMED	88	Manhattan	10,84%	1,71%	1,15%	4,54%
		Propuesta	11,36%	1,80%	1,22%	4,52%

Donde se presenta el descriptor original resaltado en amarillo, y los resultados obtenidos con cada variante de PCMSD y la variante CMED con la medida de similitud *Manhattan* y la propuesta. Se observó que algunas de las variantes superan al original, principalmente CMED que fue la variante más estable, ya que obtuvo mejores resultados en los cuatro conjuntos de imágenes. Las variantes mejoran principalmente con el

conjunto de datos Caltech-101, por lo que el uso de estructuras de diferentes tamaños ayuda en la recuperación de nivel dos, que se refiere a los objetos. Por otro lado, los cambios realizados en la cuantificación no solo contribuyeron a la reducción de la dimensión del vector de características, sino que también mejoraron el rendimiento general del algoritmo.

A pesar de que se obtienen buenos resultados en caltech-101 con las variantes que realizan un submuestreo piramidal, no se logra superar en todos los conjuntos al descriptor original, por lo que se decidió evaluar las ocho diferentes "profundidades" posibles. Buscando mejorar la recuperación al detectar la profundidad con mejores resultados, en este trabajo se utilizó la variante con mejor resultado, siendo en este caso la variante PCMSD con 298 valores. Los resultados se muestran de la Tabla 5.9 a la Tabla 5.12, donde se presentan resultados favorables en algunas profundidades que mejoran en comparación a la profundidad cuatro que es la utilizada por defecto.

Con el objetivo de mejorar los resultados y considerando que CMED demostró las mejores evaluaciones, se decidió evaluar el desempeño del descriptor propuesto utilizando el submuestreo piramidal. A diferencia de CMSD, el descriptor propuesto calcula los tipos de estructuras únicamente en la última profundidad. Los resultados obtenidos con las tres formas propuestas para obtener el vector utilizando el submuestreo piramidal con una profundidad de cuatro se muestran en la Tabla 5.13. Se puede observar que el descriptor propuesto, CMED, muestra mejores resultados con una longitud total de 298, sumados a los 27 valores del vector de estructuras.

Tabla 5.9 Evaluación de profundidades de PCMSD con Corel-1k

Profundidad	P	R	MAP	ANMRR
1	77,50%	9,30%	8,66%	1,31%
2	77,50%	9,30%	8,67%	1,31%
3	78,58%	9,43%	8,85%	1,24%
4	79,08%	9,49%	8,91%	1,21%
5	78,67%	9,44%	8,91%	1,24%
6	79,00%	9,48%	8,98%	1,22%
7	79,00%	9,48%	8,98%	1,22%
8	79,00%	9,48%	8,98%	1,22%

Tabla 5.10 Evaluación de profundidades del PCMSD con Corel-5k

Profundidad	P	R	MAP	ANMRR
1	32,32%	3,88%	2,84%	4,01%
2	32,07%	3,85%	2,88%	4,03%
3	32,27%	3,87%	2,90%	4,01%
4	32,22%	3,87%	2,89%	4,02%
5	32,22%	3,87%	2,88%	4,01%
6	32,22%	3,87%	2,88%	4,01%
7	32,22%	3,87%	2,88%	4,01%
8	32,22%	3,87%	2,88%	4,01%

Tabla 5.11 Evaluación de profundidades del PCMSD con Corel-CBIR

Profundidad	P	R	MAP	ANMRR
1	37,27%	3,99%	3,27%	1,45%
2	37,51%	4,05%	3,31%	1,43%
3	37,41%	4,03%	3,28%	1,44%
4	37,34%	4,02%	3,27%	1,44%
5	37,62%	4,05%	3,30%	1,43%
6	37,62%	4,05%	3,30%	1,43%
7	37,62%	4,05%	3,30%	1,43%
8	37,62%	4,05%	3,30%	1,43%

Tabla 5.12 Evaluación de profundidades del PCMSD con Caltech-101

Profundidad	P	R	MAP	ANMRR
1	10,13%	1,59%	1,06%	4,57%
2	10,86%	1,73%	1,17%	4,53%
3	11,16%	1,79%	1,20%	4,52%
4	11,64%	1,87%	1,23%	4,50%
5	11,93%	1,92%	1,26%	4,49%
6	12,02%	1,91%	1,29%	4,49%
7	12,01%	1,91%	1,28%	4,49%
8	12,01%	1,91%	1,28%	4,49%

Tabla 5.13 Evaluación CMED piramidal

	Longitud del vector	Corel-1k	Corel-5k	Corel-CBIR	Caltech
CMED	88+27	79,25%	32,45%	38,29%	11,97%
	352+27	79,33%	32,62%	38,84%	11,83%
	298+27	80,33%	32,83%	39,74%	12,73%

Con la finalidad de detectar qué profundidad brindaba más información sobre los tipos de estructuras, se decidió buscar evaluar la obtención de las estructuras en las tres profundidades previas. Los resultados se muestran en la Tabla 5.14, donde se observa que la profundidad cuatro entregas mejores resultados que el resto de las profundidades, sin embargo, la profundidad uno estrega resultados similares y en ocasiones mejores que la profundidad dos y tres.

Tabla 5.14 Obtención de las estructuras en diferentes profundidades

	Profundidad	Corel-1k	Corel-5k	Corel-CBIR	Caltech
CMED	1	79,33%	32,62%	38,46%	11,94%
	2	79,33%	32,55%	37,91%	11,93%
	3	79,83%	32,43%	37,42%	12,01%
	4	80,33%	32,83%	39,74%	12,73%

Asimismo, considerando que la profundidad utilizada en las variantes se estableció igual a cuatro, se evaluó el descriptor CMED con las ocho diferentes profundidades. En la Figura 5.3, se muestra los resultados obtenidos. Donde se encontró que la profundidad

cuatro es la más estable para este descriptor, con los conjuntos de imágenes utilizados en este trabajo. Se encontró que la profundidad mayor a cuatro perjudica el rendimiento del descriptor en algunos casos, lo que se relaciona a los elementos de estructuras y la cantidad de imágenes, ya que la profundidad 6 mejoraba los resultados, y los resultados disminuyen drásticamente en niveles menores con conjuntos de imágenes más grandes.

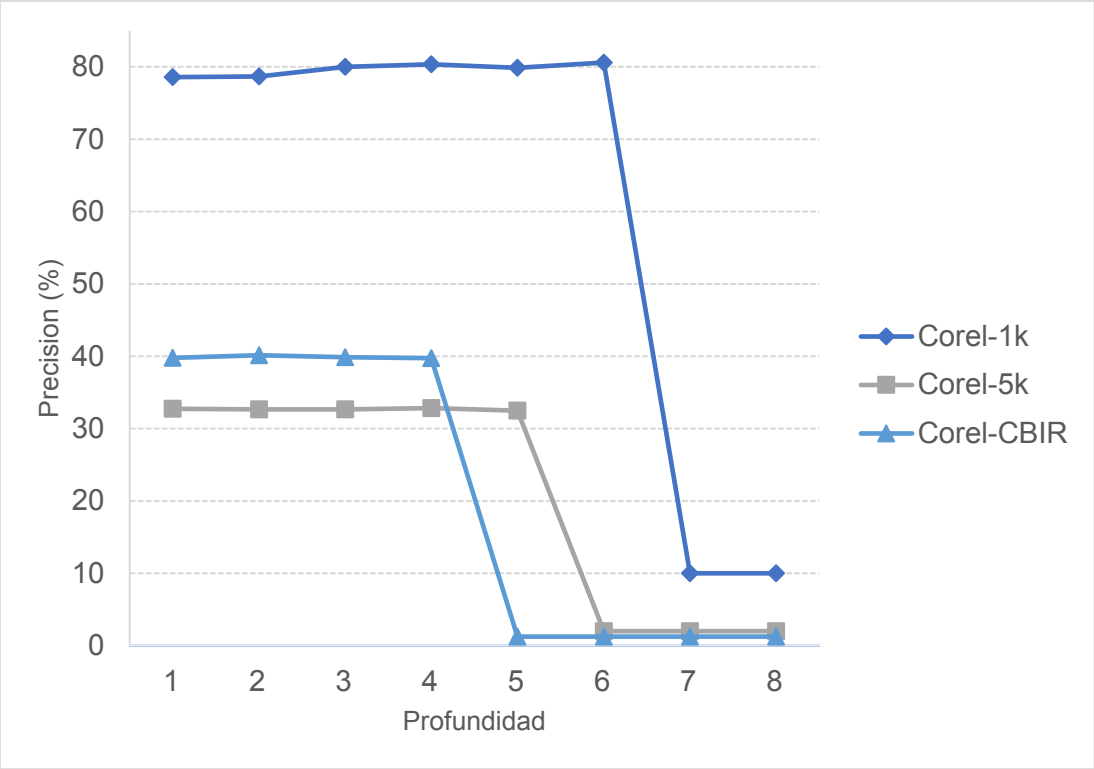


Figura 5.3 Diferentes profundidades utilizando CMED

Para encontrar la medida de similitud con el descriptor propuesto CMED, se evaluó su desempeño con el conjunto de imágenes corel-1k utilizando tres medidas de similitud. La distancia Manhattan Eq. (2.23), la distancia euclidiana Eq. (5.2) y también la medida de similitud definida para el descriptor SED Eq. (2.22). La evaluación se realizó variando el número de imágenes recuperadas K de 1 a 100. El resultado es presentado en la Figura 5.4. A pesar de que los resultados entre la distancia L_1 y la distancia propuesta para el descriptor SED L_s son muy similares, la distancia L_s presentó un mejor desempeño a lo largo del experimento. Por lo que, la medida de similitud propuesta para SED es una mejor opción que la medida de distancia L_1 normalmente utilizada para descriptores basados en microestructuras.

$$L_2(T,Q) = \sqrt{\sum_{i}^{N}(T_i - Q_i)^2} \qquad (5.2)$$

Asimismo, se evaluó el desempeño del descriptor CMED con los diferentes tipos de estructuras propuestos. La Tabla 5.15 muestra los resultados obtenidos con Corel-1k considerando las dos propuestas de estructuras, a su vez, se probó sin considerar las

estructuras de cero y ocho elementos, por lo que se evaluaron las cuatro posibles configuraciones 7, 9, 28 y 30 tipos. Además, se consideraron dos formas de obtener las estructuras: en los mapas de microestructuras (MM) y en los mapas de correlación (CM). El uso de un mayor número de tipos de estructuras mejora el rendimiento de la recuperación hasta en un 5% cuando el descriptor obtiene las estructuras en los mapas de correlación. Sin embargo, cuando se detectan las estructuras en los mapas de microestructura, el descriptor no muestra una diferencia significativa entre nueve y treinta tipos de estructuras. Por lo tanto, el uso de un mayor número de estructuras no conlleva una mejora notable en este caso.

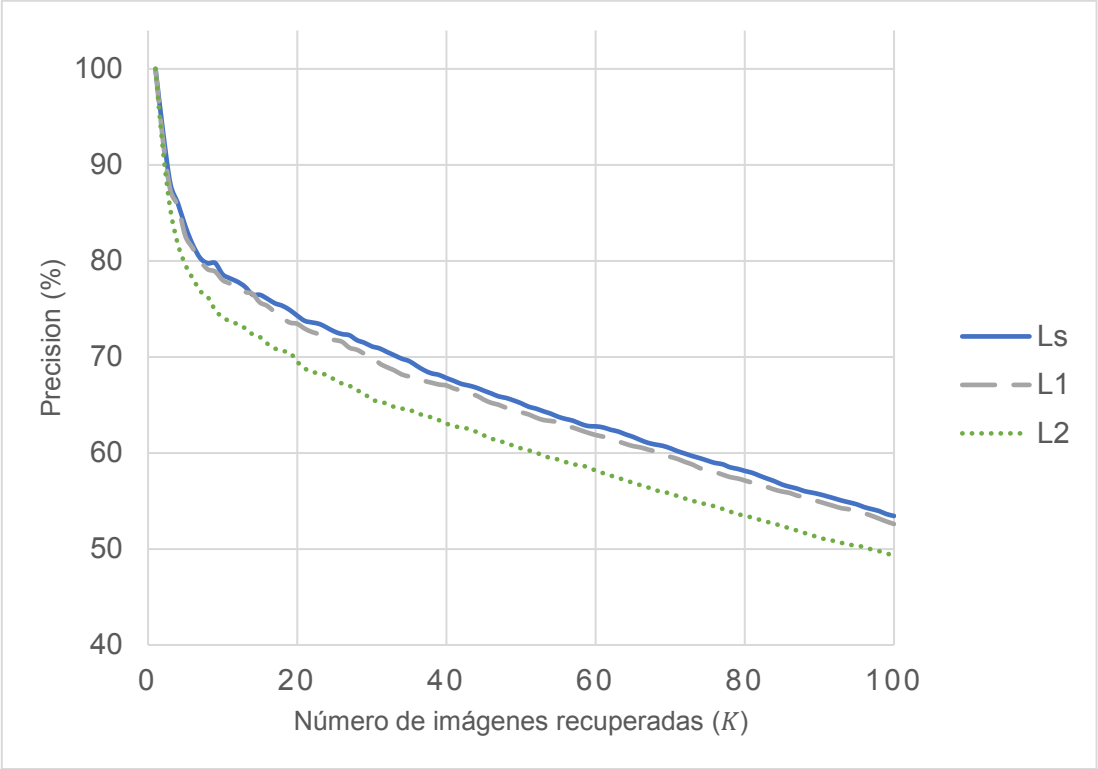


Figura 5.4 Evaluación del descriptor propuesto CMED con diferentes medidas de similitud con Corel-1k

Tabla 5.15 Resultados con diferentes formas de obtener los tipos de estructuras

Cantidad de estructuras	Obtención	P	R	MAP	ANMRR
7	MM	78.42%	9.41%	8.75%	1.25%
	CM	47.08%	5.65%	4.61%	3.12%
9	MM	78.58%	9.43%	8.78%	1.24%
	CM	44.25%	5.31%	4.31%	3.29%
28	MM	78.50%	9.42%	8.77%	1.25%
	CM	52.25%	6.27%	5.24%	2.82%
30	MM	78.58%	9.43%	8.77%	1.24%
	CM	47.83%	5.74%	4.78%	3.08%

Por otro lado, se probó con dos diferentes formas de obtener la correlación de las características, se experimentó con una correlación tipo "*AND*" y tipo "*OR*". Los resultados en la Figura 5.5, muestran que la diferencia es pequeña en comparación a los resultados obtenidos en otros experimentos, sin embargo, la correlación tipo "OR", en general se mantiene por encima de la correlación "AND", superándolo únicamente en el conjunto Caltech-101. Por otra parte, se probó el descriptor con dos diferentes corrimientos de la ventana 3×3, con $stride = 1$ y con $stride = 3$, siendo la cantidad de movimiento o corrimiento de la ventana sobre la imagen. Los resultados obtenidos fueron muy similares, sin embargo, el uso de un $stride$ mayor reduce el costo computacional por lo que, el $stride = 3$ parece ser una mejor opción para el tipo de imágenes utilizadas.

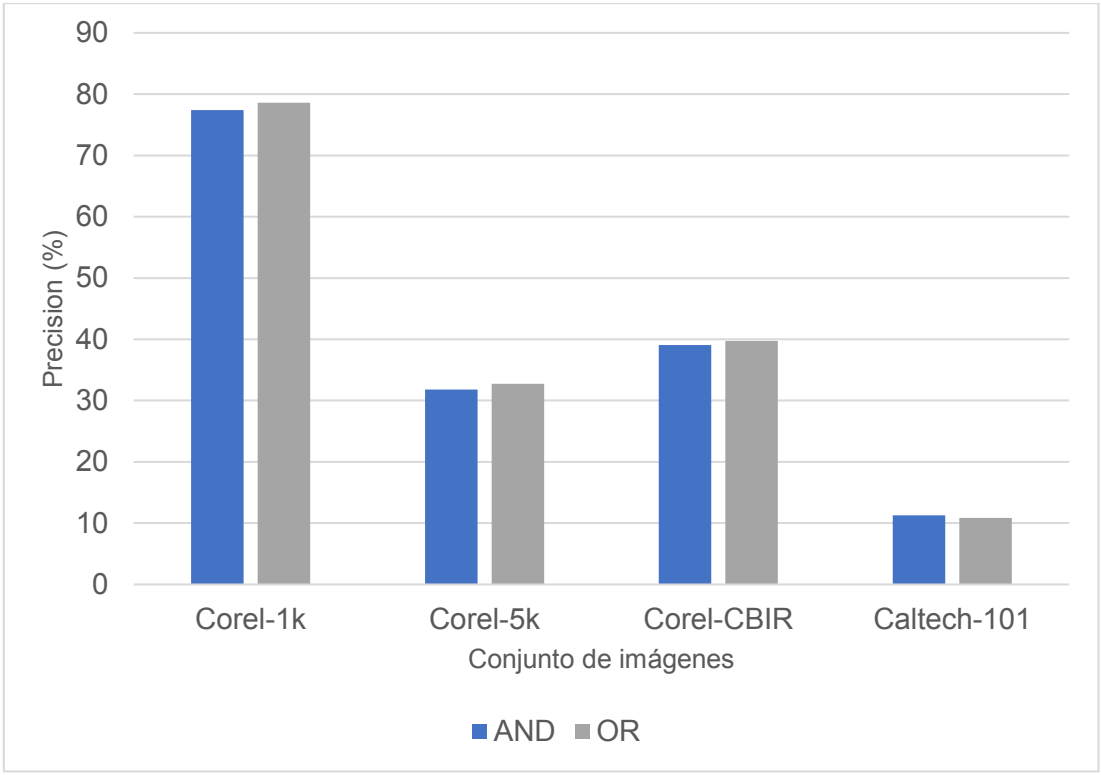


Figura 5.5 Resultados con correlación tipo "AND" y "OR"

Finalmente, a fin de mejorar el descriptor propuesto y considerando que la extracción del borde con calculo fraccionario entrega bordes representativos y tolera diferentes tipos de ruido. Se implementó y evaluó el desempeño del descriptor propuesto, utilizando cálculo fraccionario en la obtención del mapa de borde. Se utilizó con un $\alpha = 1.5$, es decir se obtuvo la una y media derivada. Los resultados de la recuperación utilizando el conjunto de imágenes Corel-1k en comparación a Sobel ordinario se muestran en la Figura 5.6. En los resultados se pude observar que a pesar de que la detección de bordes utilizando el cálculo fraccionario parece obtener buenos resultados en términos generales, en la descripción se obtienen resultados muy similares en cuanto a las métricas, sin embargo, las imágenes contenidas en los conjuntos de imágenes utilizados no poseen el tipo de ruido al que se han encontrado tolerantes, lo que podría presentar una desventaja para el enfoque. Asimismo, son propuestas muy recientes que aún falta por estudiar.

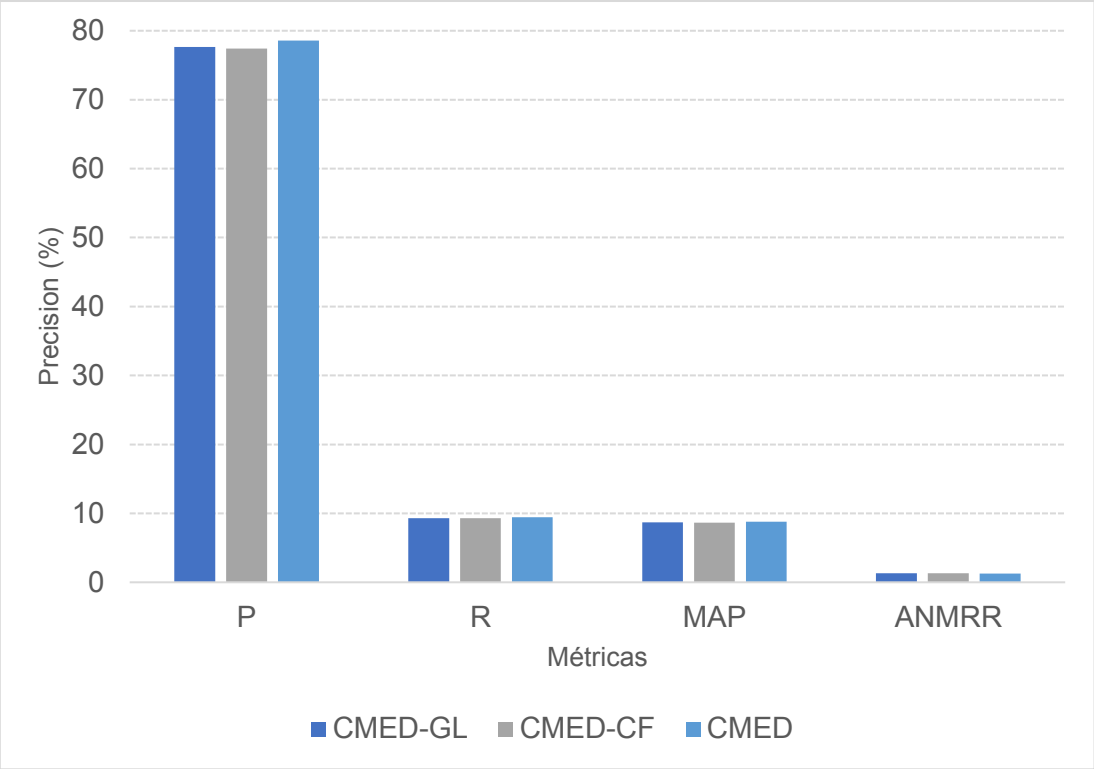


Figura 5.6 Evaluación de Sobel fraccionario

5.2.1.3 Discusión

Los experimentos mostraron que algunas de las variantes superan a los descriptores originales como MIFH-R1, PCMSD y CMED. De las variantes propuestas, se puede considerar como la mejor a CMED. En general, se logró superar el desempeño del descriptor CMSD, utilizando los elementos de estructura y utilizando diferentes tamaños de estructuras. Por lo que considerar los elementos de estructuras en la imagen mejoran la recuperación en nivel 2 y 3. Sin embargo, falta profundizar en los tipos de estructura ya que solo se experimentó con nueve elementos de estructura, en los cuales podría existir información importante ya que, pueden presentarse diferentes formas con la misma cantidad de elementos.

En promedio CMED logró mejorar la recuperación en hasta un 2.83% con *Precision*, 1.23% con *Recall*, 1.26% con *MAP* y 0.17% con *ANMRR*, en comparación con CMSD, siendo el mejor evaluado de los descriptores encontrados en la literatura. Lo que es una mejora significante considerando los resultados obtenidos en la literatura por loas diferentes propuestas en el área de recuperación de imágenes por contenido en imágenes naturales. Considerando que los conjuntos de imágenes utilizados en este trabajo requieren una recuperación en nivel 2 y 3, es decir, recuperaciones con clases semánticas, como es el caso de los conjuntos Corel-1k, Corel-5k y Corel-CBIR. El descriptor CMED y todas las variantes que superan al descriptor CMSD, logra mejorar la representación de características de alto nivel, por lo que reducen la brecha semántica en relación con los descriptores actuales.

En cuanto al cálculo fraccional no se obtuvo una mejora en los resultados de recuperación, sin embargo, el hecho de aplicar el MD-TOD utilizando un detector de

bordes fraccionario, parece prometedor al tener un detector Sobel fraccionario a color, asimismo, faltaría probar con diferentes propuestas de detectores de borde fraccionario y con imágenes contaminadas con ruido, lo que podría dar una variante de CMED con cierta tolerancia al ruido.

Finalmente, a pesar de los resultados obtenidos a lo largo de los experimentos, aún falta complementar con otros conjuntos de imágenes, al igual que relacionar los elementos de estructuras de cada característica, y probar con diferentes características de bajo nivel, ya que no se ha contemplado forma o posición espacial. Asimismo, el área requiere investigación e información, ya que no sólo podrían beneficiar a los sistemas de recuperación de imágenes, sino también a cualquier otro sistema que requiera el análisis de imágenes. Considerando que actualmente, la mayoría de los sistemas utilizan descriptores que representan una característica de bajo nivel y no consideran las relaciones entre ellos, perdiendo información relevante que no se encuentra directamente en la imagen.

5.2.2 Evaluación del desempeño del descriptor propuesto CMED

En general el mejor rendimiento se obtuvo con el descriptor propuesto CMED. En específico utilizando nueve estructuras obtenidas en los mapas de microestructuras con una relación “OR”. Por lo que se continuó con la evaluación en comparación con otros descriptores utilizados para CBIR.

5.2.2.1 Comparación del descriptor CMED con otros descriptores en la recuperación de imágenes

Para este experimento, se utilizaron los conjuntos de imágenes, Corel-1k, Corel5k y Corel-CBIR, ya que son conjuntos que requieren de un sistema nivel tres de recuperación. El descriptor propuesto CMED se comparó contra los descriptores más utilizados en el estándar MPEG-7, EHD y CLD [47]. Del mismo modo, se seleccionaron descriptores del estado del arte que utilizan microestructuras como MSD [121] y CMSD [122]. También se tomaron algunos descriptores basados en estructuras como SED [109] y basados en la integración de funciones como MIFH [95]. Teniendo en cuenta que no todos los descriptores han sido evaluados utilizando los mismos conjuntos de imágenes y que los experimentos no siempre se describen en detalle. Para dar una evaluación justa, se implementaron todos los descriptores en el mismo entorno y con las configuraciones recomendadas por los autores.

De la Figura 5.7 a la Figura 5.9 se muestran los resultados obtenidos con *Precision* en los tres conjuntos de imágenes. Los gráficos muestran que el descriptor propuesto CMED funciona mejor en todos los conjuntos de imágenes. Lo que indica que CMED consigue recuperar en promedio un mayor número de imágenes correctas que el resto de los descriptores. En las gráficas se observa que CMED tiene una menor pendiente en la reducción de la precisión cuando K aumenta, y se mantiene por encima de los demás durante todas las variaciones.

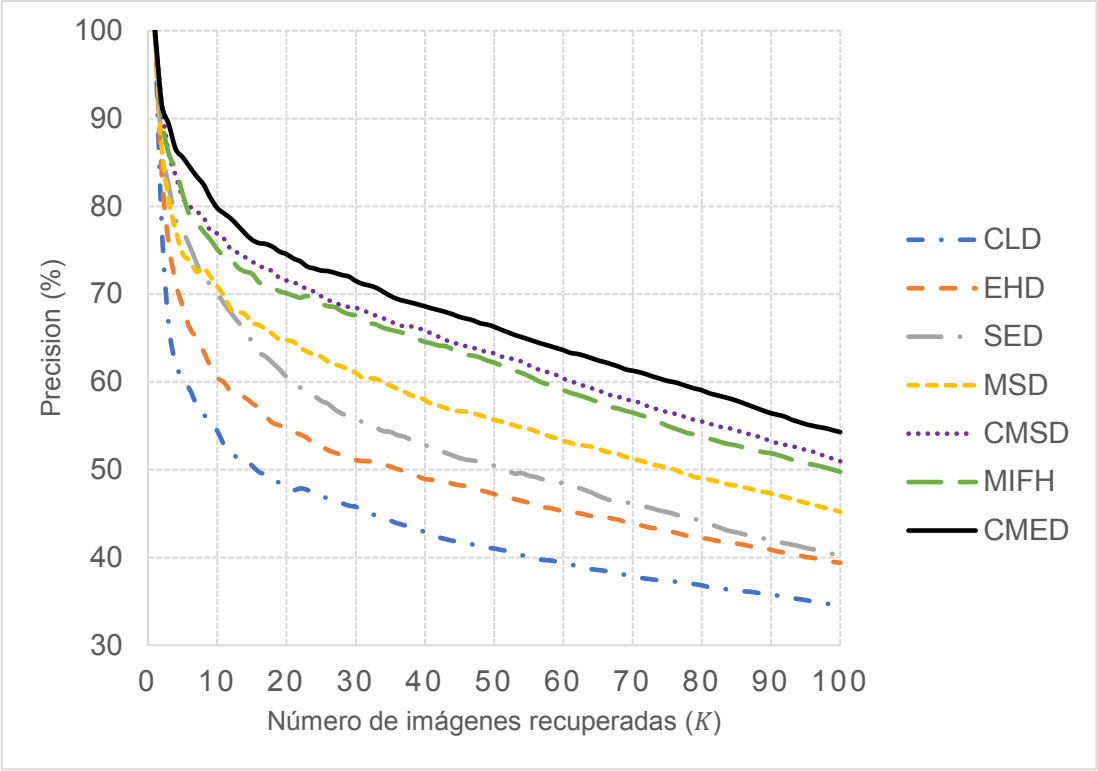


Figura 5.7 Precision de CMED en comparación con otros descriptores con Corel-1k

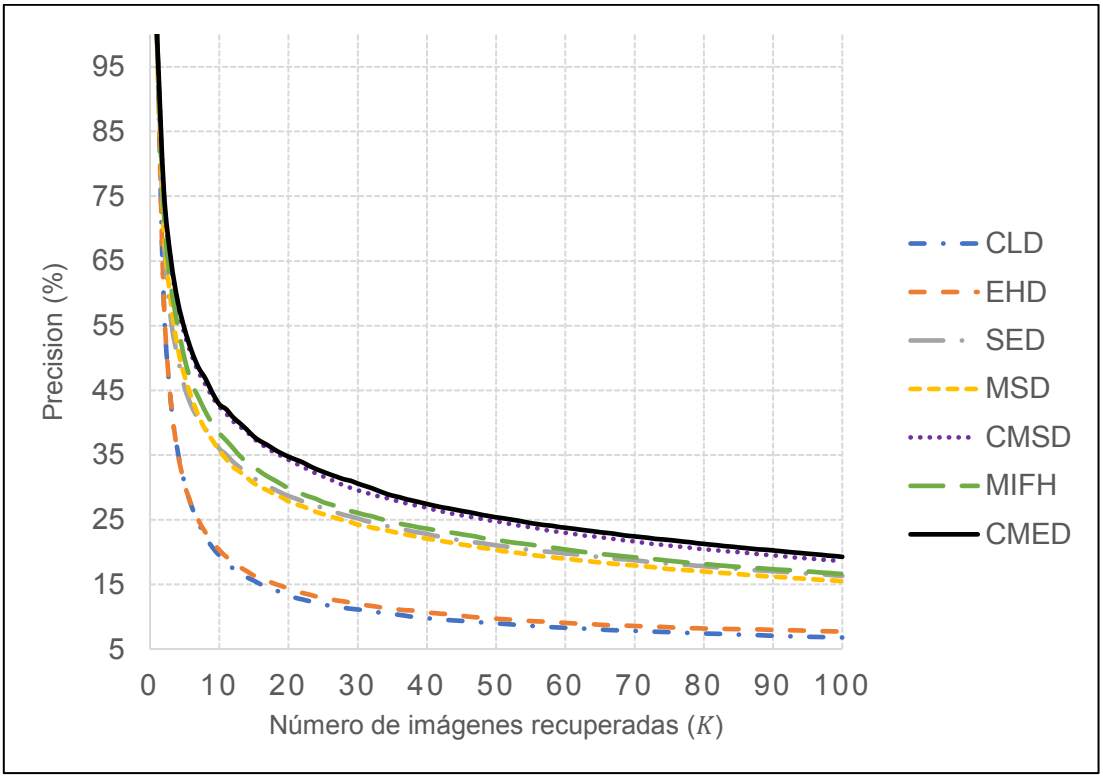


Figura 5.8 Precision de CMED en comparación con otros descriptores con Corel-5k

Cuando $K = 12$, siendo un valor comúnmente utilizado para evaluar el desempeño de un descriptor, se encontró que el descriptor propuesto CMED, obtiene un 78.58% de

precisión en Corel-1k. Esto representa una mejora por encima del 19% en relación con los descriptores propuestos por la norma y del 3,5% en relación con el descriptor CMSD, siendo el de mejor desempeño de los descriptores del estado del arte. En cambio, con Corel-5k y $K = 12$, CMED obtiene una mejora de más del 22% respecto al estándar MPEG-7, y consigue mantenerse por encima del descriptor CMSD. Con Corel-CBIR, el descriptor propuesto mejora en más de un 18% los del estándar, y por encima de un 3.5% los del estado del arte.

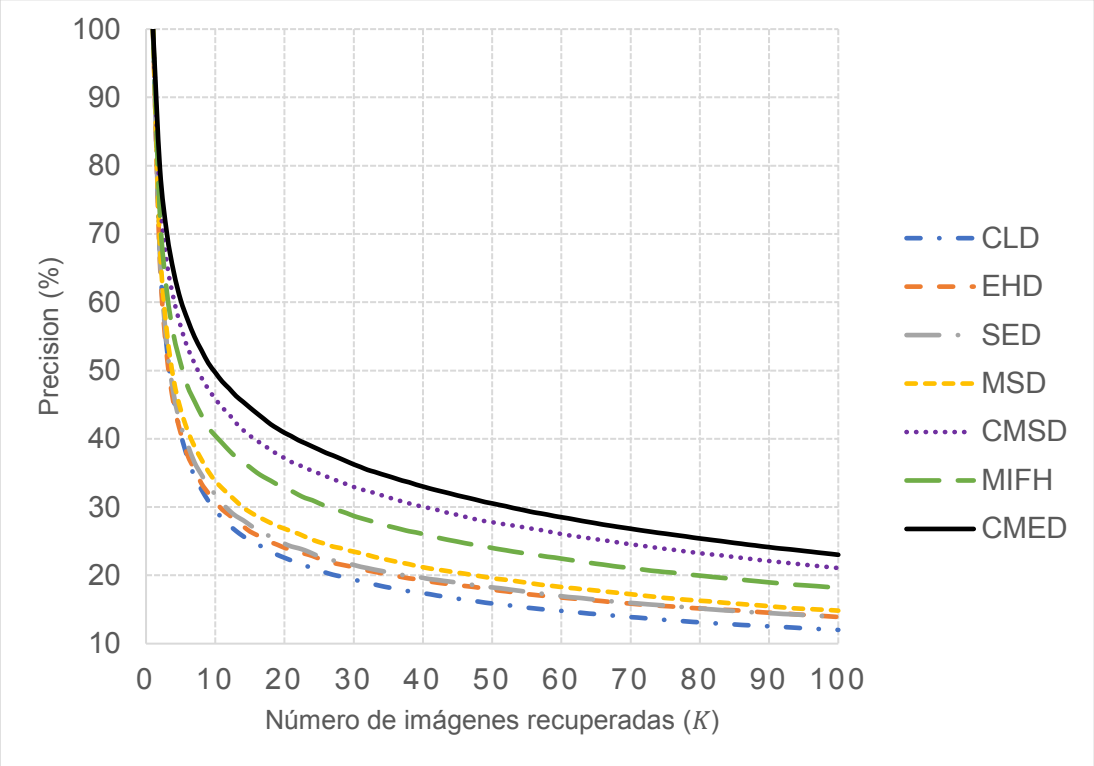


Figura 5.9 Precision de CMED en comparación con otros descriptores con Corel-CBIR

El gráfico de la Figura 5.10 muestra los resultados con la métrica de recuperación en Corel-1k. La métrica de recuperación muestra cuántas imágenes relevantes pueden recuperar el descriptor en la consulta. Una recuperación perfecta daría una recta de 45°. Ya que aumenta proporcionalmente en relación con el valor de K. En el gráfico se puede observar que el descriptor propuesto tiene un aumento más lineal que el resto de los descriptores, lo que indica que tiene un mejor rendimiento de recuperación. Por otra parte, se puede observar que el descriptor está cada vez más separado del resto de descriptores, por lo que es capaz de recuperar más imágenes que el resto de los descriptores a medida que aumenta el número de imágenes recuperadas K. Así como, cuando $K = 12$, se encontró que CMED obtiene un 9.43% de *Recall* en Corel-1k, siendo 12% una recuperación perfecta. En relación con los otros descriptores, el descriptor propuesto obtiene un mejor rendimiento en el conjunto de imágenes de Corel-1k. Por otro lado, con Corel-5k Figura 5.11 y Corel-CBIR Figura 5.12, se puede observar el mismo resultado que con Corel-1k en relación con los demás descriptores. A pesar de que con Corel-5k los resultados son muy cercanos a CMSD, con el conjunto de imágenes Corel-CBIR se puede observar una mayor separación en cada incremento en número de imágenes recuperadas K.

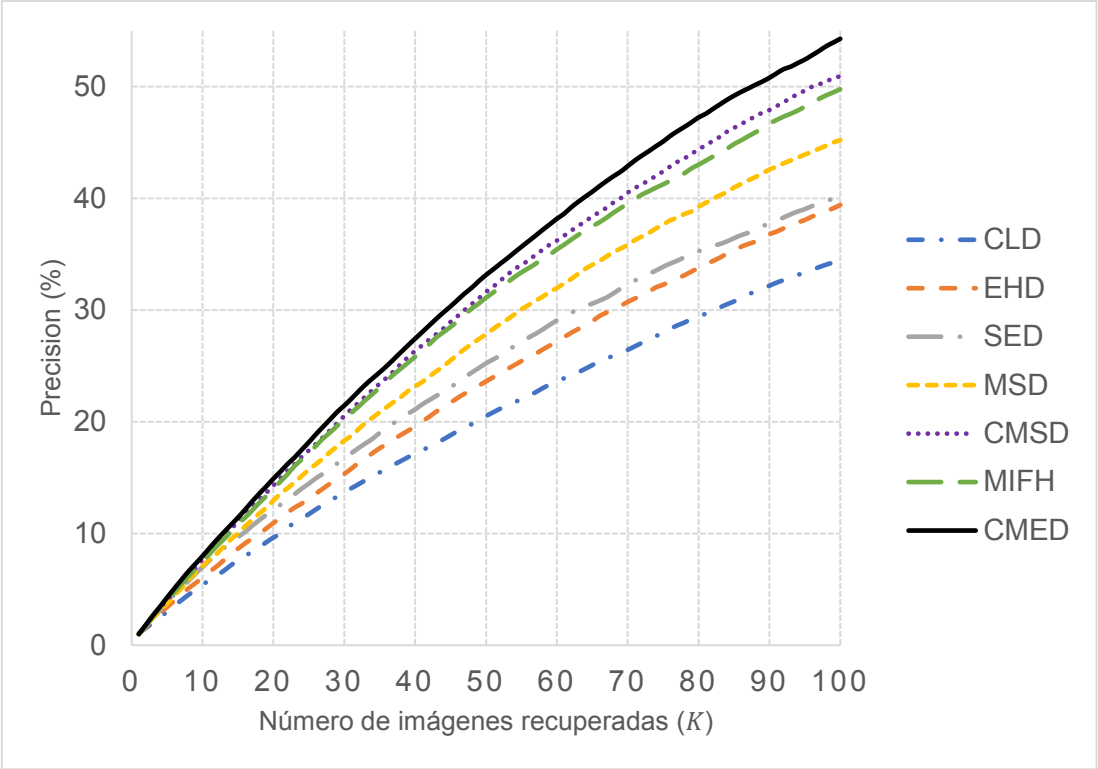


Figura 5.10 Recall de CMED en comparación con otros descriptores con Corel-1k

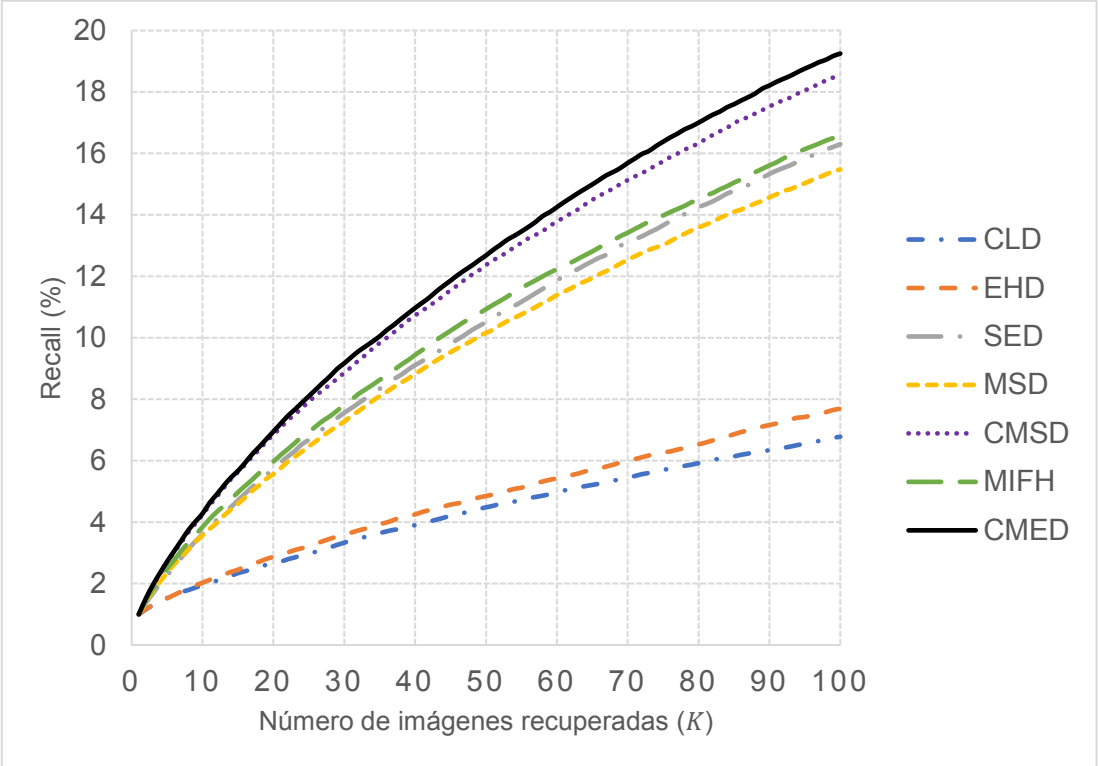


Figura 5.11 Recall de CMED en comparación con otros descriptores con Corel-5k

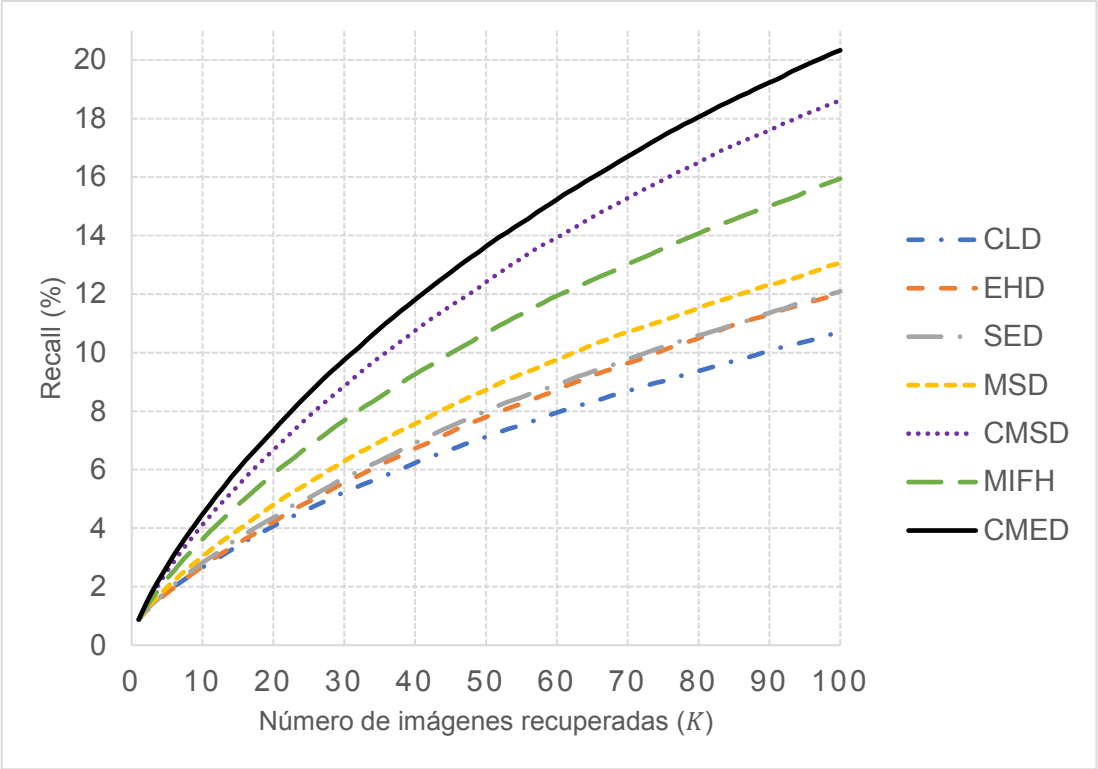


Figura 5.12 Recall de CMED en comparación con otros descriptores con Corel-CBIR

En la Figura 5.13, Figura 5.14 y Figura 5.15 se presentan los gráficos de los resultados obtenidos utilizando la métrica MAP, que evalúa tanto la *Presicion* como *Recall*, es decir, cuántas imágenes relevantes se recuperan en relación con la cantidad total de imágenes en cada categoría.

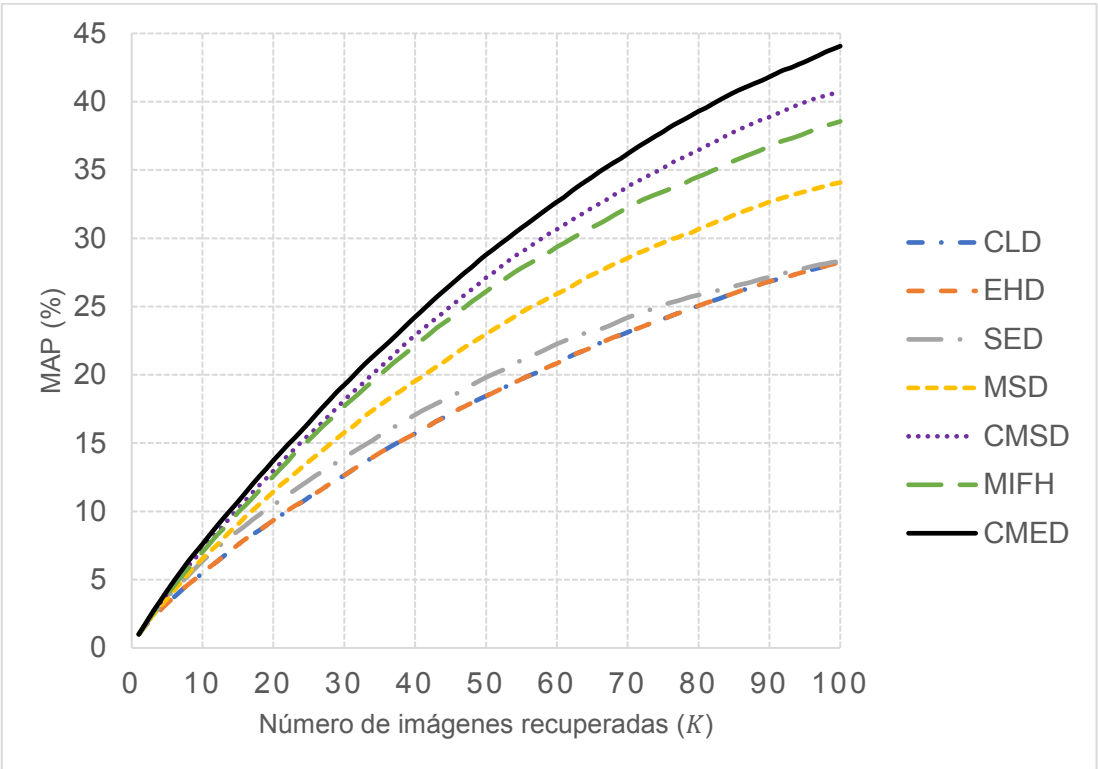


Figura 5.13 MAP de CMED en comparación con otros descriptores con Corel-1k

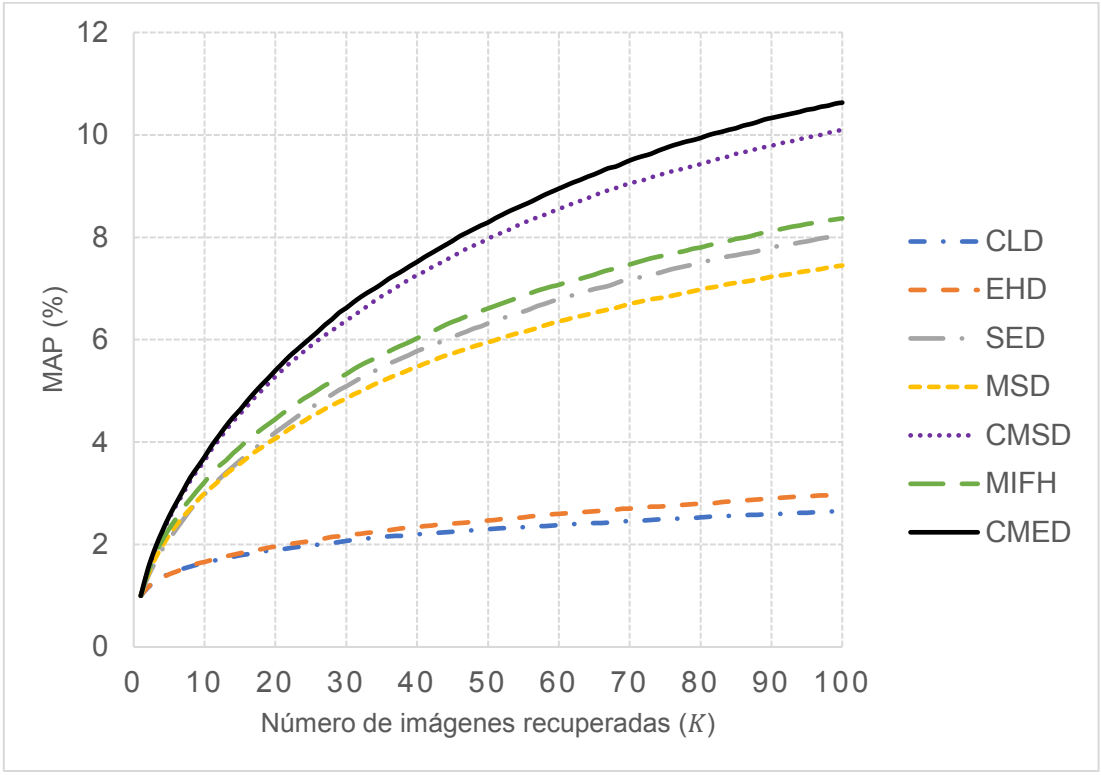


Figura 5.14 MAP de CMED en comparación con otros descriptores con Corel-5k

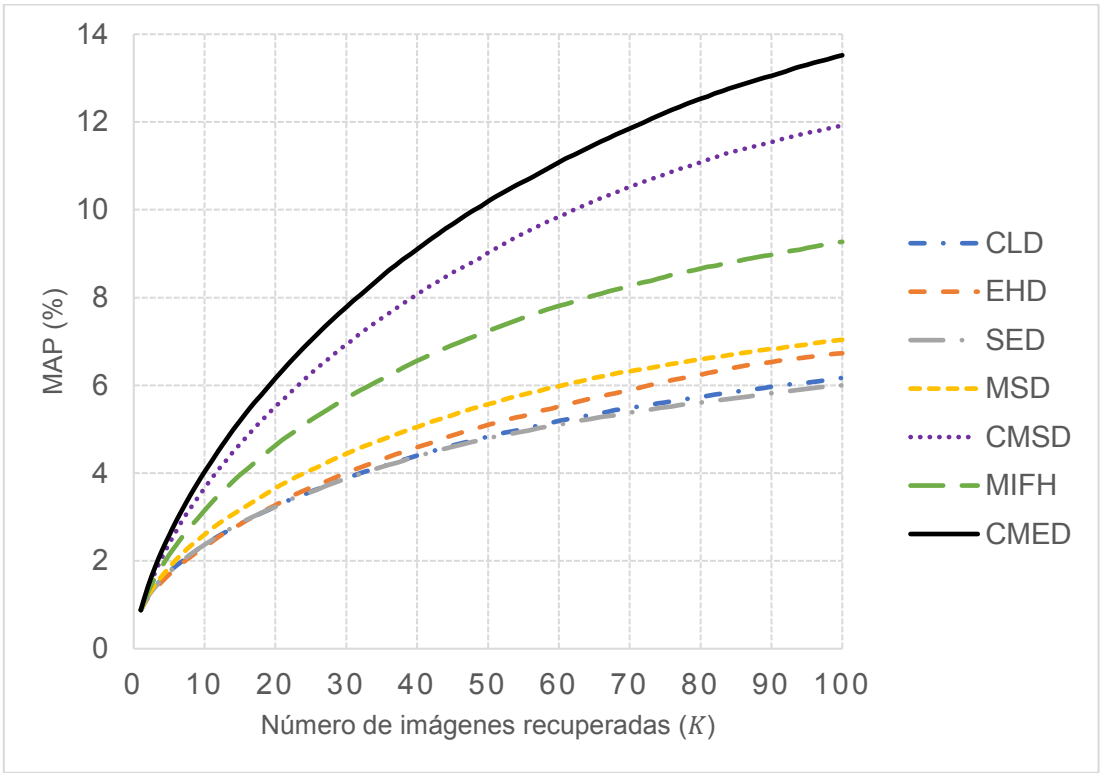


Figura 5.15 MAP de CMED en comparación con otros descriptores con Corel-CBIR

Los gráficos muestran que CMED supera en rendimiento al resto de los descriptores. Además, se observa que, al igual que con la métrica de recuerdo, el descriptor propuesto se distancia cada vez más de los demás a medida que aumenta el valor de K. Además,

cuando $K = 12$, los resultados del descriptor propuesto se mantienen consistentemente por encima de los demás descriptores.

La Figura 5.16 presenta los resultados obtenidos con la métrica *ANMRR* en los tres conjuntos de imágenes. *ANMRR* a diferencia de las demás métricas, contempla la posición en la que se recuperó cada imagen perteneciente a la clase de consulta. Por este motivo, el número de imágenes recuperadas no afecta a su resultado. *ANMRR* brinda una evaluación general ya que no solo considera cuántas imágenes recupera cada descriptor, sino que también considera las posiciones en las que fueron recuperadas. La Figura 5.16, muestra que el descriptor propuesto CMED presenta un mejor rendimiento de recuperación, obteniendo un 34.79%, lo que supone una mejora de más del 15%, respecto a los descriptores del estándar MPEG-7, y de más del 3% respecto a los descriptores del estado del arte en Corel-1k. Con los conjuntos de imágenes Corel-5k y Corel-CBIR obtuvo 75.33% y 66.93%, lo cual es una mejor evaluación que el resto de los descriptores. Los resultados con $ANMRR$ indican que el descriptor no solo es capaz de recuperar más imágenes de media que el resto de los descriptores, sino que también es capaz de recuperarlas en una mejor posición.

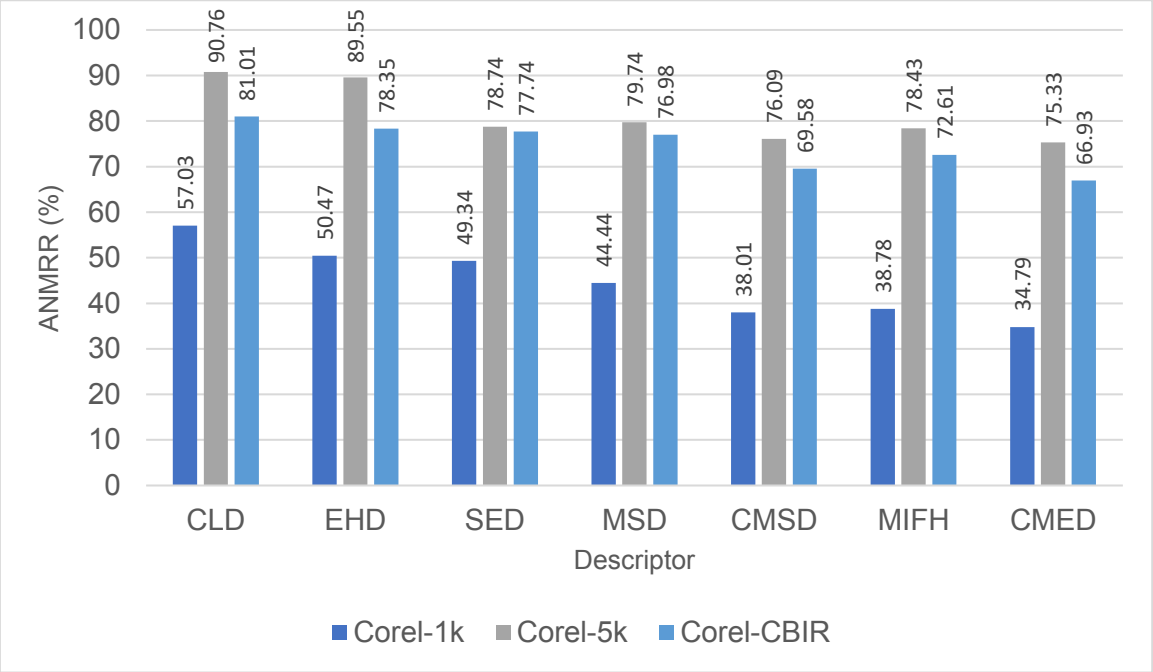


Figura 5.16 ANMRR de CMED en comparación con otros descriptores con Corel-1k, Corel-5k y Corel-CBIR

Con el fin de determinar en qué categorías se obtienen los mejores resultados y si existe algún sesgo o categoría desafiante para el descriptor, se presentan los resultados en términos de precisión para cada categoría del conjunto de datos Corel-1k con $K = 12$ en la Tabla 5.16 . Además, en la Tabla 5.17 se muestran los resultados utilizando la métrica *ANMRR*. Al analizar las tablas, se puede observar que los mejores resultados se obtienen con MIFH, CMSD y el descriptor propuesto CMED. También se encontró que el descriptor propuesto supera a los descriptores del estándar MPEG-7 en la mayoría de las categorías y, en general, muestra mejores resultados que el resto de los

descriptores en África, Playa, Autobús, Elefante, Flor, Caballo y Montaña. Sin embargo, CMED presenta un rendimiento inferior en la categoría de Edificios en comparación con los descriptores del estándar MPEG-7. Aparentemente, el descriptor responde mejor a imágenes naturales con muchos cambios de textura y color. A pesar de que el descriptor no logra superar a todos los descriptores en todas las categorías, en promedio muestra un desempeño mejor, ya que se mantiene entre los mejores. Por lo tanto, el descriptor propuesto exhibe un rendimiento más estable en comparación con el resto de los descriptores.

Tabla 5.16 Precision por categorías del descriptor CMED en contraste al resto de descriptores con Corel-1k

	CLD	EHD	SED	MSD	CMSD	MIFH	CMED
África	26.67%	35.83%	82.50%	55.00%	76.67%	77.50%	85.00%
Playa	20.00%	25.00%	65.00%	72.50%	78.33%	70.83%	77.50%
Edificio	64.17%	45.83%	31.67%	31.67%	40.83%	37.50%	50.83%
Autobús	31.67%	43.33%	55.83%	73.33%	74.17%	74.17%	75.00%
Dinosaurio	21.67%	93.33%	71.67%	80.83%	80.00%	80.00%	83.33%
Elefante	100.00%	99.17%	90.83%	100.00%	100.00%	99.17%	100.00%
Flor	46.67%	38.33%	56.67%	60.00%	73.33%	69.17%	78.33%
Caballo	75.83%	90.83%	85.00%	82.50%	86.67%	80.83%	88.33%
Montaña	98.33%	83.33%	75.83%	95.00%	95.00%	95.83%	97.50%
Comida	36.67%	33.33%	63.33%	31.67%	45.83%	55.00%	50.00%
Promedio	52.17%	58.83%	67.83%	68.25%	75.08%	74.00%	78.58%

Tabla 5.17 ANMRR por categorías del descriptor CMED en contraste al resto de descriptores con Corel-1k

	CLD	EHD	SED	MSD	CMSD	MIFH	CMED
África	75.72%	65.90%	38.37%	56.85%	46.02%	40.03%	37.49%
Playa	86.08%	73.82%	57.19%	44.31%	38.88%	48.25%	35.14%
Edificio	51.43%	66.97%	74.01%	72.18%	64.07%	62.84%	56.38%
Autobús	69.53%	73.96%	52.36%	47.28%	39.10%	39.47%	43.26%
Dinosaurio	82.45%	21.96%	41.28%	31.81%	33.49%	27.39%	28.97%
Elefante	0.55%	1.67%	23.61%	4.69%	0.77%	8.01%	0.02%
Flor	59.83%	69.61%	61.85%	54.00%	43.07%	46.45%	40.76%
Caballo	47.99%	28.60%	28.10%	35.17%	29.55%	33.64%	27.72%
Montaña	33.15%	32.30%	54.00%	34.99%	27.52%	28.79%	23.56%
Comida	63.60%	69.90%	62.67%	63.10%	57.67%	52.96%	54.55%
Promedio	57.03%	50.47%	49.34%	44.44%	38.01%	38.78%	34.79%

Por otra parte, en la descripción de imágenes se busca obtener un vector que sea representativo aún bajo diferentes transformaciones. Para evaluar si un descriptor es tolerante a las transformaciones no puede considerarse únicamente la medida de similitud, ya que no todos utilizan la misma y cada descriptor tiene una dimensión y rangos diferentes en cada valor. Por ello, se decidió evaluar el descriptor propuesto usando las mismas consultas, pero sujeto a diferentes transformaciones. Para este experimento se consideraron cinco transformaciones: rotación de 90°, rotación de 180°,

cambio de escala del 50%, cambio de escala del 80% y, finalmente, una transformación tipo espejo.

La Tabla 5.18, La Tabla 5.18 muestra los resultados obtenidos por cada descriptor con Corel-1k en las cinco transformaciones diferentes, evaluados mediante la métrica ANMRR. Por otro lado, la Tabla 5.19 presenta los resultados en términos del porcentaje de veces en que la imagen original se recupera en la primera posición. Se encontró que el descriptor propuesto logró recuperar la imagen original en primera posición en todas las consultas, lo cual indica una buena tolerancia, ya que se espera que la diferencia entre la imagen afectada y la original sea cero. Además, se observó un mejor desempeño en todas las transformaciones en comparación con el resto de los descriptores en términos de ANMRR. Considerando los resultados obtenidos con el conjunto de imágenes Corel-1k, el descriptor CMED mostró una mayor tolerancia a la rotación, la escala y la transformación espejo en comparación con los demás descriptores, ya que logró obtener un rendimiento superior y recuperar la imagen original en todas las consultas.

Tabla 5.18 Resultados con ANMRR bajo diferentes transformaciones utilizando Corel-1k

Transformación	CLD	EHD	SED	MSD	CMSD	MIFH	CMED
Escala 50%	57.03%	50.63%	49.91%	52.76%	45.60%	53.05%	40.15%
Escala 80%	57.04%	50.61%	48.35%	45.60%	38.94%	40.41%	35.91%
Rotación 180°	61.52%	56.85%	49.61%	43.75%	38.03%	38.86%	34.81%
Rotación 90°	62.87%	65.14%	52.46%	44.14%	38.40%	39.20%	35.39%
Espejo	57.40%	50.30%	49.82%	44.51%	38.01%	38.86%	34.80%
Promedio	59.17%	54.70%	50.03%	46.15%	39.80%	42.07%	36.21%

Tabla 5.19 Resultados obteniendo la imagen original en la primera posición utilizando Corel-1k

Transformación	CLD	EHD	SED	MSD	CMSD	MIFH	CMED
Escala 50%	47.00%	38.00%	98.00%	100.00%	100.00%	100.00%	100.00%
Escala 80%	86.00%	49.00%	100.00%	100.00%	100.00%	100.00%	100.00%
Rotación 180°	100.00%	100.00%	93.00%	97.00%	99.00%	75.00%	100.00%
Rotación 90°	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
Espejo	98.00%	82.00%	100.00%	100.00%	100.00%	100.00%	100.00%
Promedio	86.20%	73.80%	98.20%	99.40%	99.80%	95.00%	100.00%

Finalmente, se presenta la Tabla 5.20, donde se muestran los datos de los descriptores del estado del arte de la Tabla 3.1. Donde se especifica el tipo de características de bajo nivel utilizadas, el método que, implementado para obtener información a partir de la relación de las características, el tamaño del vector de características y el porcentaje de mejora con respecto al estándar MPEG-7. El porcentaje de mejora se estableció con respecto a la métrica ANMRR con el conjunto de imágenes Corel-1k, y considerando el descriptor del estándar MPEG-7 mejor evaluado, siendo en este caso EHD con un *ANMRR de* 50.47%.

En la Tabla 5.20 se puede observar que el descriptor propuesto presenta una mejora mayor en relación con el resto de los descriptores. Por otro lado, a pesar de que el descriptor propuesto CMED obtiene un vector mayor que algunos de los descriptores, logra mantenerse dentro del rango de los descriptores utilizando por el estándar ya que

CLD obtiene un vector de 192, lo que supone una reducción de 77 valores. En general, el descriptor CMED logra considerar e integrar diferentes métodos sin aumentar, e incluso disminuyendo en algunos casos el vector de características en comparación al estándar MPEG-7 y mejorando la representación en las imágenes en comparación al resto de descriptores.

Tabla 5.20 Descriptores del estado del arte implementados y CMED, en contraste al estándar MPEG-7

Descriptor	Características utilizadas	Método utilizado	Longitud del vector	Mejora en relación con EHD
MSD	• Color • Borde	Microestructuras	72	6.03%
SED	• Color	Estructuras	360	1.13%
CMSD	• Color • Borde • Intensidad	Microestructuras	88	12.46%
MIFH	• Color • Borde	Integración de características	112	11.69%
CMED	• Color • Borde • Intensidad	Microestructuras, Estructuras, integración de características	115	15.68%

5.2.2.2 Comparación CMED con características profundas

Considerando el número de trabajos encontrados en el estado del arte que utilizan características profundas para la descripción de imágenes, se realizaron experimentos en comparación con el descriptor propuesto CMED. Para los experimentos se utilizaron Corel-1k, Corel-5k y Corel-CBIR. Se comparó el descriptor con dos diferentes arquitecturas de redes neuronales para la obtención de las características profundas, utilizando las cuatro métricas. La metodología realizada para obtener los vectores de características profundas con cada una de las arquitecturas se presenta a continuación, seguido de los experimentos y resultados obtenidos con cada uno de los conjuntos de imágenes.

Para los experimentos se utilizó *VGG-16*, siendo esta arquitectura frecuentemente utilizada en recuperación de imágenes. Se tomaron todas las capas de convolución, así como la primera capa completamente conectada, la Figura 5.17 muestra de manera gráfica la parte de obtención. La arquitectura se obtuvo con *Pythorch* del módulo *Torchvision*, y se utilizó el modelo pre entrenado con *ImageNet*, el cual es un conjunto de imágenes con mil clases, por lo que se obtienen características profundas para diferentes tipos de imágenes.

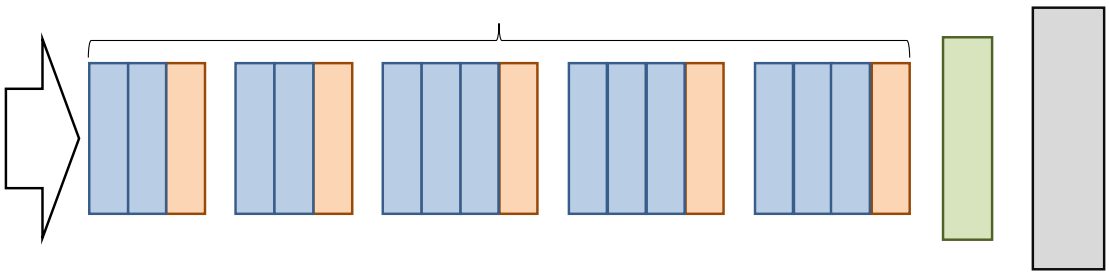


Figura 5.17 Obtención del vector de características con VGG-16 [63]

El vector resultante de *VGG-16* después del aplanamiento de los mapas de características es de 4,096 valores. Por otro parte, se realizó una normalización utilizando la función *linalg.norm(),* de la paquetería *Numpy*, siguiendo la Eq. (5.3). Para medir la distancia entre dos imágenes se utilizó la distancia *Manhattan* Eq. (2.23).

$$NorVector = \frac{Vector}{linalg.norm(Vector)} \tag{5.3}$$

Para los experimentos se utilizó *ResNet18* y al igual que *VGG-16* se obtuvo del módulo de *Torchvision* y se utilizó pre entenado con *ImageNet*. Se utilizaron todas las capas de convolución y la capa donde se realiza el *Average Pooling*, seguido del aplanamiento de los mapas, el cual se tomó como vector de características para realizar la recuperación. La Figura 5.18, muestra la parte de la arquitectura donde se extraen las características profundas. El uso de *ResNet18* da como resultado un vector de 512 características. Los valores del vector obtenido se normalizaron de la misma forma que *VGG-16* Eq. (5.3). Y para medir la distancia entre dos vectores se utilizó la Eq. (2.23).

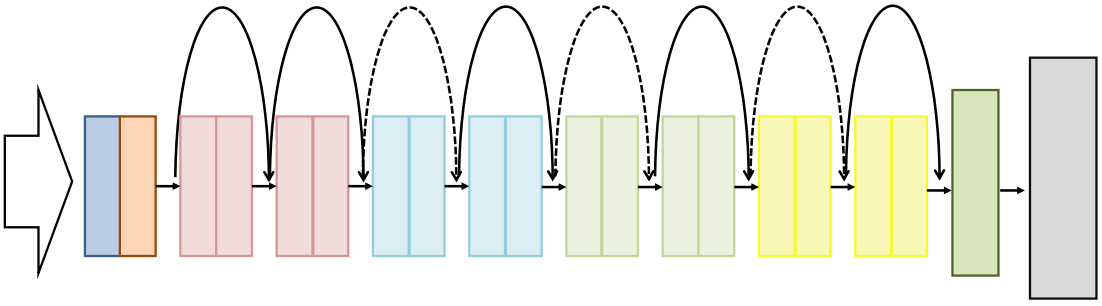


Figura 5.18 Obtención de vector de características con ResNet18 [64]

Para el experimento con Corel-1k se tomaron 10 imágenes aleatorias como consulta por clase, dando un total de 100 consultas. Considerando que las personas buscan tener un resultado en las primeras imágenes recuperadas se estableció el sistema en K = 12. Las consultas se utilizaron para cada arquitectura y el descriptor *CMED*. Los resultados obtenidos con cada una de las métricas se muestran en la Figura 5.19. Los resultados muestran que el descriptor obtenido con *VGG-16* obtiene los mejores resultados

superando a *CMED* hasta en 10% con *Precision*, sin embargo, *ResNet*, no logra superarlo, obteniendo resultados inferiores hasta en un 20%.

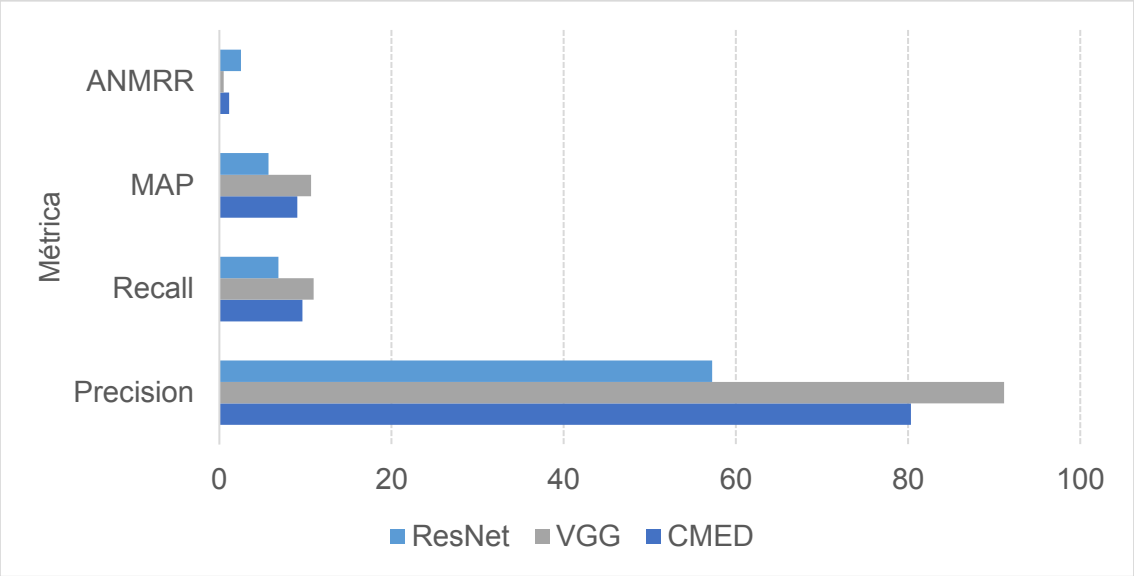


Figura 5.19 Desempeño de CMED en comparación con características profundas con Corel-1k

Con el conjunto de imágenes Corel-5k se consideró la misma cantidad de imágenes recuperadas K = 12 y las 10 consultas por clase dando un total de 500 consultas por descriptor. La Figura 5.20, muestra los resultados obtenidos, donde se pueden observar valores considerablemente menores en comparación a corel-1k, siendo un conjunto con imágenes con contornos negros que introduce ruido y con mayor cantidad de clases e imágenes, sin embargo, *VGG*-16 se mantiene por encima de los descriptores obteniendo una mejor recuperación.

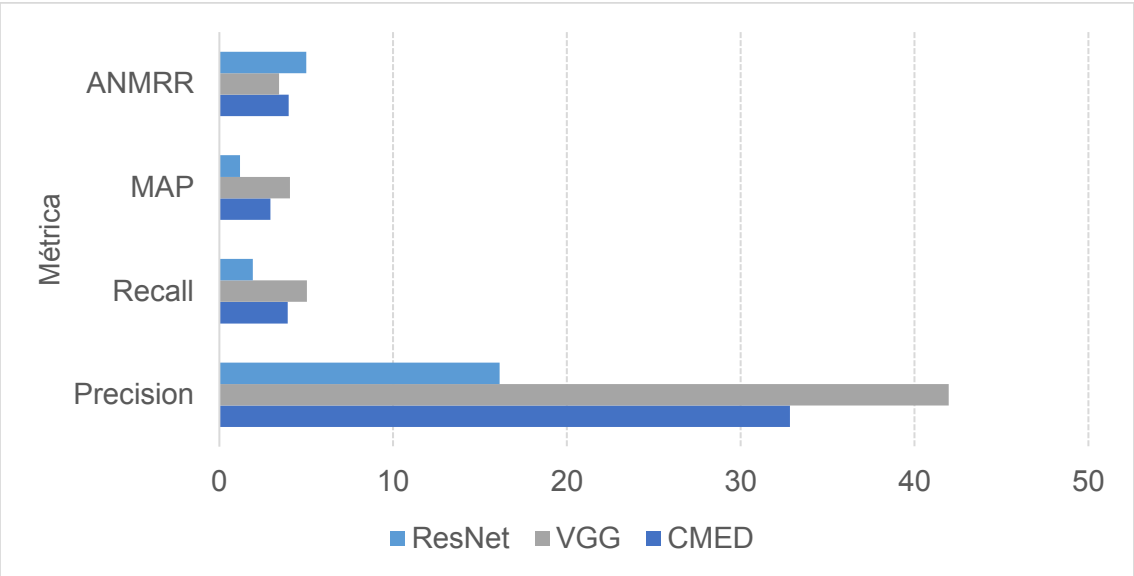


Figura 5.20 Desempeño de CMED en comparación con características profundas con Corel-5k

Para el conjunto de imágenes Corel-CBIR que contiene 10,800 imágenes con 80 clases desbalanceadas, se tomaron 10 imágenes aleatorias de cada clase con el mismo K = 12. Con lo que se realizaron 800 consultas para cada descriptor. Los resultados

obtenidos con cada una de las métricas se muestran en la Figura 5.21. Los resultados muestran que *VGG*-16 logra hasta un 30% de mejora con *Precision* en comparación a *CMED*, siendo el conjunto de imágenes donde se encontró la diferencia más grande.

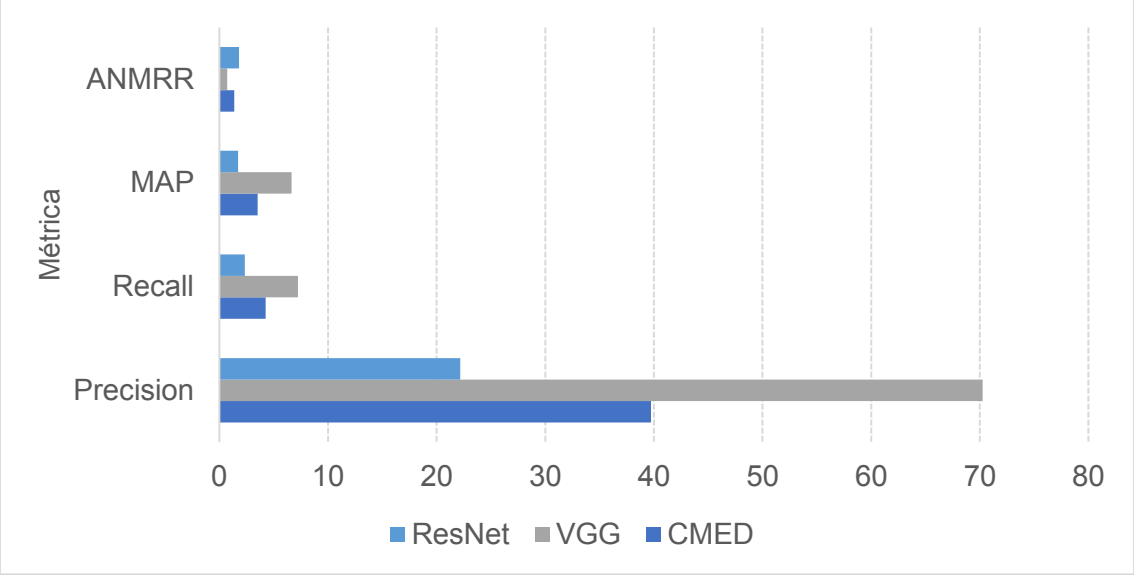


Figura 5.21 Desempeño de CMED en comparación con características profundas con Corel-CBIR

Los resultados muestran que *VGG-16* obtiene mejores resultados en los tres conjuntos de imágenes, obteniendo un resultado superior hasta por 30% en *Precision*, 3% en *Recall*, 3% con *MAP* y 0.5% con *ANMRR*. Sin embargo, *ResNet* no logra superar al descriptor propuesto en ninguno de los conjuntos, a pesar de tener 17 capas, siendo mayor que *VGG*. Aparentemente el hecho de ocupar redes con arquitectura residual para obtener el vector no presenta una ventaja en la recuperación de imágenes naturales, sin embargo, faltaría evaluar el desempeño de otras configuraciones con *ResNet*.

De manera general, el uso de características profundas obtenidas con *VGG-16* mejora la recuperación, sin embargo, el descriptor genera un vector con una longitud de 4,096 valores y requiere de obtener más de 2,000 mapas de características a diferencia de CMED que genera un vector de 115 valores utilizando un total de nueve mapas. De manera gráfica en la Figura 5.22, se presenta la *Precision* obtenida en relación con la longitud del vector. La Figura 5.23, presenta la *Precision* con respecto a la cantidad de mapas que requieren generar, lo que puede traducirse en costo computacional y tiempo de procesamiento.

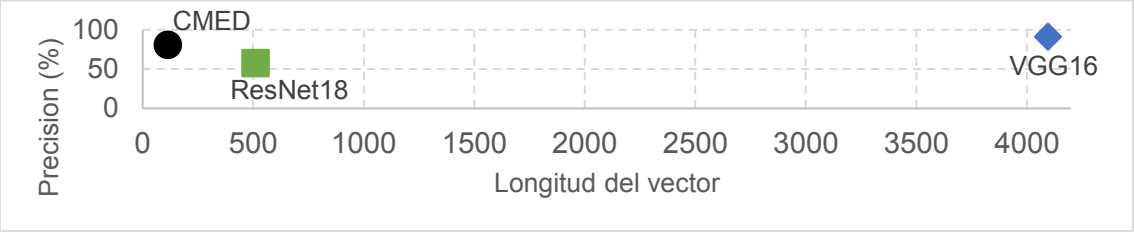


Figura 5.22 Relación Precision - longitud con CMED y las características profundas

Figura 5.23 Relación Precision - mapas con CMED y las características profundas

5.3 Análisis de los resultados

En general, la configuración óptima del descriptor propuesto CMED es con una profundidad de cuatro niveles. Sin embargo, esta configuración requiere un mayor costo computacional aproximadamente 43% mayor, considerando el consto requerido para un nivel uno de profundidad. Por otro lado, el uso de más tipos de estructuras no mejora la representación de los conjuntos de imágenes, sino que aumenta el costo computacional y la longitud del descriptor. Así mismo, el descriptor propuesto supera a otras técnicas en conjuntos que requieren recuperación de nivel tres, como son SED, CMTH, MIFH, CMSD, MSD. La diferencia entre CMED y las demás técnicas es similar a la diferencia entre CMSD y MSD en términos de significancia.

Las condiciones y la cantidad de imágenes por clase en el conjunto afectan el rendimiento de los descriptores en términos de las métricas. Las pequeñas diferencias se deben principalmente a la cantidad de imágenes que se consideraron en la recuperación. Se encontró que el descriptor propuesto no solo mejora en una categoría específica, sino que mantiene un rendimiento estable en la gran mayoría de ellas, obteniendo un mejor desempeño en promedio. Además, CMED presenta un mejor rendimiento de recuperación cuando la imagen de consulta se somete a rotación, escala y transformación en espejo. El uso de estructuras y microestructuras de elementos permite una mejor recuperación de imágenes naturales en clases semánticas en comparación con los descriptores del estándar MPEG-7 y los encontrados en el estado del arte.

En comparación con las características profundas, se encontró que el aprendizaje profundo mejora la recuperación de imágenes naturales en clases semánticas. Sin embargo, los vectores de características resultantes son considerablemente grandes, lo que dificulta la indexación del vector en los datos de contenido multimedia. Además, el uso de todas las capas convolucionales puede no ser lo suficientemente general, ya que las características obtenidas en el último nivel pueden ser muy específicas para el tipo de imágenes con las que se entrenaron.

Las microestructuras presentan ventajas como no requerir escalamiento de la imagen de consulta ni entrenamiento, así como tener vectores de menor longitud. Sin embargo, es necesario analizar qué tipos de clases ofrecen mejores resultados con las características profundas y verificar si están relacionadas con las clases presentes en el conjunto de imágenes ImageNet. También sería necesario comparar el rendimiento en relación con la longitud del vector y el tiempo de procesamiento de la imagen. En este aspecto, las microestructuras podrían presentar una ventaja al utilizar solo tres

mapas de características en comparación con las redes neuronales convolucionales, que emplean más de 64 mapas en cada capa de convolución.

Por otro lado, las principales desventajas de las características profundas en comparación con los descriptores de microestructuras son el tamaño del vector y el costo computacional para obtenerlo. En cuanto a las microestructuras, dependen de las características de bajo nivel utilizadas. Si se obtienen características poco representativas o correlacionadas, la detección de microestructuras podría resultar en una descripción poco discriminante. Teniendo en cuenta las limitaciones de cada enfoque, el uso de un enfoque híbrido que combine microestructuras y características profundas podría ofrecer mejores resultados. Al detectar las microestructuras en las características profundas, se podría obtener un vector de características más discriminante, incluso utilizando menos capas de convolución y un vector de menor longitud en comparación con el obtenido únicamente con características profundas.

6 Conclusión

Después de realizar las evaluaciones y experimentos correspondientes, se puede concluir que es posible mejorar el rendimiento obtenido por los descriptores estándar del MPEG-7 en la recuperación de imágenes en clases semánticas mediante la relación entre las características de bajo nivel y sus estructuras. En particular, el descriptor propuesto CMED logra mejorar la representación de las características de alto nivel en las imágenes mediante la relación entre las características de bajo nivel y sus estructuras. Además, se han obtenido mejoras significativas en la recuperación de nivel tres en comparación con el estado del arte.

Así mismo, se llevaron a cabo experimentos utilizando técnicas no tradicionales, como el cálculo fraccionario, y se evaluaron los resultados en comparación con las características profundas, las cuales no estaban consideradas en ellos objetivos y alcances del proyecto. Sin embargo, estos experimentos revelaron las ventajas del descriptor propuesto, ya que no requiere entrenamiento, tiene un menor costo computacional y genera vectores de menor tamaño que facilitan su indexación. En consecuencia, se logró cumplir plenamente con los objetivos, alcances y actividades establecidos en el proyecto, enriqueciendo la investigación con actividades adicionales.

No obstante, se identificaron problemáticas en los conjuntos de imágenes y en las métricas de evaluación. Se observó que existe una única métrica específica para evaluar los sistemas de recuperación, lo que limita la capacidad de evaluar los resultados en la recuperación y la capacidad de evaluar la relación semántica entre las imágenes. Es crucial desarrollar métricas o conjuntos de imágenes adicionales que consideren la relación semántica, ya que la recuperación de imágenes relacionadas en un concepto semántico que pertenecen a diferentes clases se consideran un error de la misma magnitud que imágenes que no tienen ninguna relación semántica entre ellas. En resumen, se ha demostrado el potencial de mejora en la recuperación de imágenes semánticas, pero aún se requiere un desarrollo de métricas y conjuntos de imágenes para una evaluación precisa.

Los productos generados y las principales contribuciones de la investigación se presentan a continuación. Asimismo, se destaca el cumplimiento de los objetivos, alcances y actividades establecidas en la propuesta inicial.

6.1 Productos Académicos

Durante el desarrollo de la investigación se han obtenido los siguientes productos:

1. Resumen para el Consorcio del *Mexican International Conference on Artificial Intelligence* 2019 (MICAI 2019) (Anexo A), donde se presentó el resumen de la propuesta del proyecto
2. Artículo para la 4° Jornada de Ciencia y Tecnología Aplicada (4° JCYTA) (Anexo B), donde se presenta un breve análisis a los descriptores del estándar
3. Artículo para el congreso mexicano *International Conference on Mechatronics, Electronics and Automotive Engineering* 2020 (ICMEAE 2020) (Anexo C), donde se presenta el análisis de los descriptores del estándar para CBIR
4. Artículo para la 5° Jornada de Ciencia y Tecnología Aplicada (5° JCYTA) (Anexo D), donde se presenta el análisis del estado del arte sobre la problemática y obtención de características de alto nivel
5. Presentación de divulgación en la Universidad Tecnológica Emiliano Zapata para el "*2022 IEEE School on Computational Intelligence and Robotics"* llamada "Recuperación de imágenes mediante descriptores MPEG-7"
6. Artículo en la revista indexada en JCR *Traitement du signal* (TS) Vol. 39 No. 1 (Anexo E), donde se presenta el análisis para la detección de debilidades y algunas propuestas de mejora para subsanarlas
7. Artículo enviado para la revista JCR *Journal of Visual Communication and Image Representation* (JVCI)-23-521, donde se presenta el descriptor propuesto CMED
8. Reportes y presentaciones semestrales con avances y resultados parciales del proyecto
9. Documento y presentación de examen predoctoral donde se presenta la propuesta de solución
10. Redacción y entrega de tesis con los avances y resultados completos del proyecto de investigación

6.2 Aportaciones

Las aportaciones realizadas hasta el momento son:

- Estudio del estado del arte de la recuperación de imágenes basada en contenido.
- Implementación de los dos descriptores del estándar MPEG-7 más utilizados para CBIR.
- Implementación de cuatro descriptores encontrados en el estado del arte.
- Modificación al código del descriptor CMSD proporcionado por los autores.
- Estudio de variantes de los descriptores MIFH y CMSD.

- Propuestas e implementación de estructuras basado en cantidad de elementos.
- Propuesta e implementación para la detección de estructuras elementales tolerantes a rotación y efecto espejo.
- Propuesta e implementación de submuestreo piramidal para CMSD y CMED.
- Propuesta e implementación del descriptor CMED para la recuperación de imágenes basado en contenido que supera los descriptores actuales.
- Estudio para obtención de mejor configuración para el descriptor propuesto CMED en imágenes naturales.
- Detección de orientación de borde en imágenes a color utilizando cálculo fraccionario.
- Implementación y comparación de características profundas en la recuperación de imágenes naturales.

6.3 Cumplimiento de los objetivos, alcances y actividades

Los objetivos, alcances y actividades establecidas para el proyecto de investigación se muestran en las Tabla 6.1, Tabla 6.2. y Tabla 6.3, respectivamente.

Tabla 6.1 Cumplimiento de los objetivos

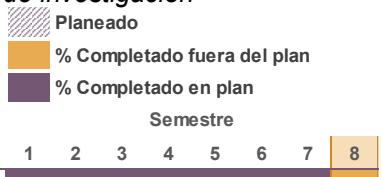
Objetivo	Porcentaje Completado	Actividades Realizadas	Productos
Estudiar los descriptores actuales.	100%	Búsqueda y análisis del estado del arte sobre descriptores MPEG-7 y nuevas propuestas para CBIR	Resumen para MIACAI 2019; Artículo en 4° JCYTA; Artículo en ICMEAE 2020; y Artículo en 5° JCYTA
Detectar deficiencias.	100%	Análisis para detectar las debilidades	Artículo JCR en TS
Proponer mejoras	100%	Propuestas para MIFH y CMSD.	Artículo JCR en TS; y Artículo JCR enviado
Aplicarlo a un dominio aceptado por la comunidad internacional como plataforma de prueba	100%	No se detectó plataforma estándar. Se realizaron pruebas sobre conjuntos de imágenes y métricas comúnmente utilizadas	Artículo JCR en TS; y Artículo JCR enviado

Tabla 6.2 Cumplimiento de los alcances

Alcance	Porcentaje Completado	Actividades Realizadas
El descriptor mejorado deberá superar alguna deficiencia relevante encontrada en descriptores anteriores.	100%	El descriptor propuesto CMED obtiene una mejor representación en términos semánticos y en tolerancia a transformaciones en comparación con el estado del arte y el estándar MPEG-7.
Las pruebas se harán sobre un dominio aceptado internacionalmente y controlado, cuyos resultados correctos sean conocidos de antemano.	100%	Los experimentos se realizaron con conjuntos de imágenes, donde se conoce previamente cuales son las imágenes que el modelo debería recuperar.
Las pruebas se harán además sobre un dominio del mundo real que tenga impacto en la línea de investigación del grupo de trabajo de Inteligencia Artificial del CENIDET.	100%	Las pruebas se realizaron en conjuntos de imágenes naturales con clases semánticas, que incluyen personas, objetos y lugares.

Tabla 6.3 Cronograma de actividades del proyecto de investigación

ACTIVIDAD	INICIO	DURACIÓN	INICIO REAL	DURACIÓN REAL	COMPLETADO
Revisión del estado del arte	1	7	1	8	100%
Estudio de algoritmos clásicos y recientes	1	3	1	3	100%
Detección de oportunidades de mejora	2	2	2	3	100%
Modelación a nivel prototipo	3	1	3	1	100%
Obtención de casos de prueba	3	1	3	1	100%
Artículo para revista mexicana	2	1	2	2	100%
Experimentos	4	3	4	4	100%
Refinación de prototipos	5	1	5	1	100%
Obtención de mejor modelo	5	2	5	2	100%
Artículo	4	1	3	1	100%
Estudio comparativo	6	1	6	1	100%
Reporte de resultados y hallazgos	6	1	6	1	100%
Artículo de resultados	6	1	5	2	100%
Estancia académica	7	1	7	1	100%
Documentación de avances y escritura de la tesis	1	8	1	8	100%



Referencias

[1] M. Sadeghi, P. Chilana, J. Yap, P. Tschandl, and M. S. Atkins, “Using content-based image retrieval of dermoscopic images for interpretation and education: A pilot study,” *Skin Research and Technology*, vol. 26, no. 4, pp. 503–512, Jul. 2020, doi: 10.1111/SRT.12822.

[2] C. G. Sotomayor *et al.*, “Content-Based Medical Image Retrieval and Intelligent Interactive Visual Browser for Medical Education, Research and Care,” *Diagnostics 2021, Vol. 11, Page 1470*, vol. 11, no. 8, p. 1470, Aug. 2021, doi: 10.3390/DIAGNOSTICS11081470.

[3] J. Tang and S. T. Acton, “A decentralized image retrieval system for education,” *Proceedings of the 2003 IEEE Systems and Information Engineering Design Symposium, SIEDS 2003*, pp. 7–12, 2003, doi: 10.1109/SIEDS.2003.157997.

[4] J. Tang and S. T. Acton, “A decentralized image retrieval system for education,” in *Syst. Inf. Eng. Design Symp.*, 2003, pp. 7–12.

[5] L. R. Nair, K. Subramaniam, and G. K. D. Prasannavenkatesan, “A review on multiple approaches to medical image retrieval system,” *Advances in Intelligent Systems and Computing*, vol. 1125, pp. 501–509, 2020, doi: 10.1007/978-981-15-2780-7_55/COVER.

[6] N. A. M. Zin, R. Yusof, S. A. Lashari, A. Mustapha, N. Senan, and R. Ibrahim, “Content-Based Image Retrieval in Medical Domain: A Review,” *J Phys Conf Ser*, vol. 1019, no. 1, 2018, doi: 10.1088/1742-6596/1019/1/012044.

[7] A. Kumar *et al.*, “Adapting content-based image retrieval techniques for the semantic annotation of medical images,” *Computerized Medical Imaging and Graphics*, vol. 49, pp. 37–45, 2016, doi: 10.1016/j.compmedimag.2016.01.001.

[8] P. Shamna, V. K. Govindan, and K. A. Abdul Nazeer, “Content-based medical image retrieval by spatial matching of visual words,” *Journal of King Saud University - Computer and Information Sciences*, vol. 34, no. 2, pp. 58–71, Feb. 2022, doi: 10.1016/J.JKSUCI.2018.10.002.

[9] S. Veerashetty and N. B. Patil, “Manhattan distance-based histogram of oriented gradients for content-based medical image retrieval,” *International Journal of Computers and Applications*, vol. 43, no. 9, pp. 924–930, 2021, doi: 10.1080/1206212X.2019.1653011.

[10] N. Darapureddy, N. Karatapu, and T. K. Battula, “Local Derivative Vector Pattern: Hybrid Pattern for Content-Based Medical Image Retrieval,” *Review of Computer Engineering Studies*, vol. 7, no. 4, pp. 79–86, Dec. 2020, doi: 10.18280/RCES.070401.

[11] M. Gaikwad and O. Hoeber, “An interactive image retrieval approach to searching for images on social media,” *CHIIR 2019 - Proceedings of the 2019 Conference on Human Information Interaction and Retrieval*, pp. 173–181, Mar. 2019, doi: 10.1145/3295750.3298930.

[12] M. S. Dao, P. Quang Nhat Minh, A. Kasem, and M. S. Haja Nazmudeen, “A context-aware late-fusion approach for disaster image retrieval from social media,” *ICMR 2018 - Proceedings of the 2018 ACM International Conference on Multimedia Retrieval*, pp. 266–273, Jun. 2018, doi: 10.1145/3206025.3206047.

[13] J. Ahmad, M. Sajjad, S. Rho, and S. W. Baik, “Multi-scale local structure patterns histogram for describing visual contents in social image retrieval systems,” *Multimed Tools Appl*, vol. 75, no. 20, pp. 12669–12692, 2016, doi: 10.1007/s11042-016-3436-9.

[14] S. V Bhoir and S. Patil, “A review on recent advances in content-based image retrieval used in image search engine,” *Library Philosophy and Practice*, pp. 1–45, 2021.

[15] D. Agrawal, A. Agarwal, and D. K. Sharma, “Content-Based Image Retrieval (CBIR): A Review,” *Lecture Notes in Electrical Engineering*, vol. 855, pp. 439–452, 2022, doi: 10.1007/978-981-16-8892-8_33/COVER.

[16] J. Assfalg, A. D. Bimbo, and P. Pala, “Using Multiple examples for content-based retrieval,” in *International Conference Multimedia and Expo*, 2000.

[17] R. S. Patil and A. J. Agrawal, “Content-based Image Retrieval Systems: A Survey,” *Advances in Computational Sciences and Technology*, vol. 10, no. 9, pp. 2773–2788, 2017, doi: 10.1109/2.410145.

[18] M. B. I. Mohamed, “A Survey on Content-based Image Retrieval,” *International Journal of Advanced Computer Science and Applications*, vol. 22, no. 4, pp. 569–578, 2017, doi: 10.3724/SP.J.1089.2010.10502.

[19] X. Li, J. Yang, and J. Ma, “Recent developments of content-based image retrieval (CBIR),” *Neurocomputing*, vol. 452, pp. 675–689, Sep. 2021, doi: 10.1016/J.NEUCOM.2020.07.139.

[20] P. Manisha, R. Jayadevan, and V. S. Sheeba, “Content-based image retrieval through semantic image segmentation,” in *AIP Conference Proceedings*, 2020, pp. 030008–1–030008–7. doi: 10.1063/5.0004087.

[21] A. Alzu’bi, A. Amira, and N. Ramzan, “Semantic content-based image retrieval: A comprehensive study,” *J Vis Commun Image Represent*, vol. 32, no. July, pp. 20–54, 2015, doi: 10.1016/j.jvcir.2015.07.012.

[22] S. M. Chavda and G. M. Mahesh, “Content-Based Image Retrieval : The State of the,” *International Journal of Next-Generation Computing*, vol. 10, no. 3, 2019.

[23] W. Zhou, H. Li, and Q. Tian, “Recent Advance in Content-based Image Retrieval: A Literature Survey,” *arXiv preprint arXiv:1706.06064*, p. 1, 2017, [Online]. Available: http://arxiv.org/abs/1706.06064

[24] M. K. Alsmadi, “Content-Based Image Retrieval Using Color, Shape and Texture Descriptors and Features,” *Arab J Sci Eng*, vol. 45, no. 4, pp. 3317–3330, Apr. 2020, doi: 10.1007/s13369-020-04384-y.

[25] X. Zhang, C. Bai, and K. Kpalma, “OMCBIR: Offline mobile content-based image retrieval with lightweight CNN optimization,” *Displays*, vol. 76, p. 102355, Jan. 2023, doi: 10.1016/J.DISPLA.2022.102355.

[26] N. F. Zulkurnain, M. A. Azhar, and M. A. Mallik, “Content-Based Image Retrieval System Using Fuzzy Colour and Local Binary Pattern with Apache Lucene,” in *Proceedings of Second International Conference on Advances in Computer Engineering and Communication Systems*, B. V. and M. R. R. and S. R. K. Reddy A. Brahmananda and Kiranmayee, Ed., Singapore: Springer Nature Singapore, 2022, pp. 13–20.

[27] Google, “Google imágenes.” https://www.google.com/imghp?hl=es

[28] T. Sikora, “The MPEG-7 visual standard for content description-an overview,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 11, no. 6, 2001, Accessed: Apr. 18, 2023. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/927422/

[29] V. Tyagi, *Content- Based Image Retrieval (Ideas, Influences, and Current Trends)*. Springer, Singapore, 2017.
[30] Y. Chen, J. Z. Wang, and R. Krovetz, “An unsupervised learning approach to content-based image retrieval,” in *IEEE Proceedings of the International Symposium on Signal Processing and its Applications,* 2003, pp. 197–200.
[31] J. Eakins and M. Graham, “Content-based image retrieval,” *Technical Report, University of Northumbria at Newcastle,*. 1999.
[32] M. W. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain, “Content-based image retrieval at the end of the early years.,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 12, pp. 1349–1380, 2000.
[33] C. Tsuhan, “From Low-Level Features to High-Level Semantics: Are We Bridging the Gap?,” *Seventh IEEE International Symposium on Multimedia (ISM’05)*, vol. 2005, pp. 179–179, Dec. 2005, doi: 10.1109/ISM.2005.62.
[34] Y. Liu, D. Zhang, G. Lu, and W. Y. Ma, “A survey of content-based image retrieval with high-level semantics,” *Pattern Recognit*, vol. 40, no. 1, pp. 262–282, 2007, doi: 10.1016/j.patcog.2006.04.045.
[35] D. G. Lowe, “Object recognition from local scale-invariant features,” *Proceedings of the IEEE International Conference on Computer Vision*, vol. 2, pp. 1150–1157, 1999, doi: 10.1109/ICCV.1999.790410.
[36] F. Taheri, K. Rahbar, and P. Salimi, “Effective features in content-based image retrieval from a combination of low-level features and deep Boltzmann machine,” *Multimed Tools Appl*, pp. 1–24, Aug. 2022, doi: 10.1007/S11042-022-13670-W/TABLES/12.
[37] N. Upadhyaya and M. Dixit, “A Review: Relating Low Level Features to High Level Semantics in CBIR,” *International Journal of Signal Processing, Image Processing and Pattern Recognition*, vol. 9, no. 3, pp. 433–444, 2016, doi: 10.14257/ijsip.2016.9.3.37.
[38] P. Sundara Vadivel, D. Yuvaraj, S. Navaneetha Krishnan, and S. R. Mathusudhanan, “An efficient CBIR system based on color histogram, edge, and texture features,” *Concurr Comput*, vol. 31, no. 12, 2019, doi: 10.1002/cpe.4994.
[39] Y. H. Lee and S. Il Bang, “Improved image retrieval and classification with combined invariant features and color descriptor,” *J Ambient Intell Humaniz Comput*, vol. 10, no. 6, pp. 2255–2264, 2018, doi: 10.1007/s12652-018-0817-0.
[40] S. Gandhani and N. Singhal, “Content Based Image Retrieval: Survey and Comparison of CBIR System based on Combined Features,” *International Journal of Signal Processing, Image Processing and Pattern Recognition*, vol. 8, no. 11, pp. 417–422, 2015, doi: 10.14257/ijsip.2015.8.11.37.
[41] Y. D. Kenchappa and K. Kwadiki, “Content-based image retrieval using integrated features and multi-subspace randomization and collaboration,” *International Journal of System Assurance Engineering and Management*, vol. 13, no. 5, pp. 2540–2550, Oct. 2022, doi: 10.1007/S13198-022-01663-9/TABLES/3.
[42] S. Fadaei, R. Amirfattahi, and M. R. Ahmadzadeh, “New content-based image retrieval system based on optimised integration of DCD, wavelet and curvelet features,” *IET Image Process*, vol. 11, no. 2, pp. 89–98, 2017, doi: 10.1049/iet-ipr.2016.0542.
[43] D. Srivastava, R. Wadhvani, and M. Gyanchandani, “A Review: Color Feature Extraction Methods for Content Based Image Retrieval,” *International Journal of Computational Engineering & Management*, vol. 18, no. 3, pp. 9–13, 2015.

[44] S. Wang, K. Han, and J. Jin, “Review of image low-level feature extraction methods for content-based image retrieval,” *Sensor Review*, vol. 39, no. 6, pp. 783–809, 2019, doi: 10.1108/SR-04-2019-0092.

[45] V. Kottawar, N. Deshpande, V. S. Jatti, and S. Bhoite, “Review Of Feature Extraction Techniques In Content Based Image Retrieval,” *J Pharm Negat Results*, vol. 13, pp. 7508–7514, Dec. 2022, doi: 10.47750/PNR.2022.13.S07.905.

[46] M. S. Lotfabadi, Y. Zhan, and A. B. Tabrizi, “A review of wrapper feature selection in content based image retrieval systems,” *Proceedings of the 2018 10th International Conference on Machine Learning and Computing*, pp. 178–183, 2018, doi: 10.1145/3195106.3195113.

[47] B. S. Manjunath, P. Salembier, and T. Sikora, *Introduction to MPEG-7. Multimedia Content Description Interface*. 2002.

[48] K. S. Aguilar-Domínguez, R. Pinto-Elías, J. G. González-Serna, and A. Magadán-Salazar, “Image Description Using the Relation Between Color and Texture in Retrieval Task,” *Traitement du Signal*, vol. 39, no. 1, pp. 21–29, Feb. 2022, doi: 10.18280/ts.390103.

[49] A. Treisman, “A feature in integration theory of attention,” *Cogn Psychol*, vol. 12, no. 1, pp. 97–136, 1980.

[50] B. Julesz, “Textons, the elements of texture perception, and their interactions.,” *Nature*, vol. 290, no. 5802, pp. 91–97, 1981, doi: https://doi.org/10.1038/290091a0.

[51] B. Julesz and J. Bergen, “Human factors and behavioral science: Textons, the fundamental elements in preattentive vision and perception of textures,” *Readings in computer vision*, vol. Elsevier, pp. 243–256, 1987.

[52] S. R. Dubey, “A Decade Survey of Content Based Image Retrieval Using Deep Learning,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 5, pp. 2687–2704, May 2022, doi: 10.1109/TCSVT.2021.3080920.

[53] R. R. Saritha, V. Paul, and P. G. Kumar, “Content based image retrieval using deep learning process,” *Cluster Comput*, vol. 22, no. 2, pp. 4187–4200, Mar. 2019, doi: 10.1007/S10586-018-1731-0/FIGURES/16.

[54] M. Tzelepi and A. Tefas, “Deep convolutional learning for Content Based Image Retrieval,” *Neurocomputing*, vol. 275, pp. 2467–2478, 2018, doi: 10.1016/j.neucom.2017.11.022.

[55] W. Chen *et al.*, “Deep Learning for Instance Retrieval: A Survey,” *IEEE Trans Pattern Anal Mach Intell*, 2022, doi: 10.1109/TPAMI.2022.3218591.

[56] M. Lux and O. Marques, *Visual information retrieval using java and lire*. 2013.

[57] C. Manning, “An introduction to information retrieval,” 2009, Accessed: Apr. 18, 2023. [Online]. Available: https://ds.amu.edu.et/xmlui/bitstream/handle/123456789/14697/Book%20558%20pages.pdf?sequence=1&isAllowed=y

[58] M. Alkhawlani, M. Elmogy, H. E. B.-Int. J. Comput. Inf. Technol, and undefined 2015, “Text-based, content-based, and semantic-based image retrievals: a survey,” *researchgate.net*, pp. 2279–0764, 2015, Accessed: Apr. 18, 2023. [Online]. Available: https://www.researchgate.net/profile/Mohammed-Elmogy/publication/273258916_Text-based_Content-based_and_Semantic-based_Image_Retrievals_A_Survey/links/54fcab360cf20700c5e96db6/Text-based-Content-based-and-Semantic-based-Image-Retrievals-A-Survey.pdf

[59] S. Singh, K. H.-I. J. of Computer, and undefined 2012, “Content based Image Retrieval based on the integration of Color histogram, Color Moment and Gabor Texture,” *Citeseer*, vol. 59, no. 17, pp. 975–8887, 2012, Accessed: Apr. 18, 2023.

[Online]. Available: https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=f3a107aefe8f78b94a6757301c8c911c6254254a

[60] R. Datta, D. Joshi, J. Li, and J. Z. Wang, “Image retrieval,” *ACM Computing Surveys (CSUR)*, vol. 40, no. 2, May 2008, doi: 10.1145/1348246.1348248.

[61] S. A. Chatzichristofis and Y. S. Boutalis, *Compact Composite Descriptors for Content Based Image Retrieval: Basics, Concepts, Tools*. VDM Verlag, 2011.

[62] B. Farnham, S. Tokyo, B. Boston, F. Sebastopol, and T. Beijing, “Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow Concepts, Tools, and Techniques to Build Intelligent Systems SECOND EDITION,” 2019.

[63] K. Simonyan and A. Zisserman, “Very Deep Convolutional Networks for Large-Scale Image Recognition,” *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, Sep. 2014, doi: 10.48550/arxiv.1409.1556.

[64] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2016-December, pp. 770–778, Dec. 2016, doi: 10.1109/CVPR.2016.90.

[65] J. E. Solís Pérez, “Cálculo Fraccionario Aplicado en Procesamiento de Imágenes y Señales,” 2020.

[66] M. Rodríguez Martín, “Introducción al cálculo fraccionario y a los modelos de crecimiento tumoral clásicos y fraccionarios,” 2019.

[67] H. Laurent, “Sur le calcul des dérivées à indices quelconques,” *Nouvelles annales de mathématiques*, 1884, Accessed: May 07, 2023. [Online]. Available: http://www.numdam.org/item/NAM_1884_3_3__240_0.pdf

[68] A. Grunwald, “Uber" begrente" Derivationen und deren Anwedung,” *Zangew Math und Phys*, 1867, Accessed: May 07, 2023. [Online]. Available: https://cir.nii.ac.jp/crid/1571980074368689536

[69] A. Letnikov, “Theory of differentiation with an arbtraly indicator,” *Matem Sbornik*, 1868, Accessed: May 07, 2023. [Online]. Available: https://cir.nii.ac.jp/crid/1573105974275528320

[70] P. A. Troncoso Rey, “Indexado y Recuperación de Imágenes por Contenido,” Tesis de maestría en ciencias , CENIDET, Cuernavaca, Morelos, 2007.

[71] C. Pérez Lara, “Recuperación Automatizada de Imágenes mediante la Implementación de Descriptores del Estándar MPEG-7,” Tesis de maestría en ciencias , CENIDET, Cuernavaca, Morelos, 2014.

[72] B. Jia, B. Meng, W. Zhang, and J. Liu, “Query rewriting and semantic annotation in semantic-based image retrieval under heterogeneous ontologies of big data,” *Traitement du Signal*, vol. 37, no. 1, pp. 101–105, 2020, doi: 10.18280/TS.370113.

[73] D. Podder, J. Mukherjee, S. M. Aswatha, J. Mukherjee, and S. Sural, *Ontology-Driven Content-Based Retrieval of Heritage Images*. 2018. doi: 10.1007/978-981-10-7221-5.

[74] F. Qin, S. Gao, X. Yang, M. Li, and J. Bai, “An ontology-based semantic retrieval approach for heterogeneous 3D CAD models,” *Advanced Engineering Informatics*, vol. 30, no. 4, pp. 751–768, 2016, doi: 10.1016/j.aei.2016.10.001.

[75] V. Vijayarajan, M. Dinakaran, P. Tejaswin, and M. Lohani, “A generic framework for ontology-based information retrieval and image retrieval in web data,” *Human-centric Computing and Information Sciences*, vol. 6, no. 1, pp. 1–30, 2016, doi: 10.1186/s13673-016-0074-1.

[76] M. Liaqat, S. Khan, and M. Majid, “Image retrieval based on fuzzy ontology,” *Multimed Tools Appl*, vol. 76, no. 21, pp. 22623–22645, 2017, doi: 10.1007/s11042-017-4812-9.

[77] V. Mezaris, I. Kompatsiaris, and M.G. Strintzis, “An ontology approach to object-based image retrieval,” in *ICIP*, 2003, pp. 511–514.

[78] G. Deepak and J. Sheeba Priyadarshini, *A hybrid semantic algorithm for web image retrieval incorporating ontology classification and user-driven query expansion*, vol. 645. Springer Singapore, 2018. doi: 10.1007/978-981-10-7200-0_4.

[79] G. Deepak and J. S. Priyadarshini, “Personalized and Enhanced Hybridized Semantic Algorithm for web image retrieval incorporating ontology classification, strategic query expansion, and content-based analysis,” *Computers and Electrical Engineering*, vol. 72, pp. 14–25, 2018, doi: 10.1016/j.compeleceng.2018.08.020.

[80] Md. F. Sadique and S. M. R. Haque, “Content-Based Image Retrieval Using Color Layout Descriptor, Gray-Level Co-Occurrence Matrix and K-Nearest Neighbors,” *International Journal of Information Technology and Computer Science*, vol. 12, no. 3, pp. 19–25, 2020, doi: 10.5815/ijitcs.2020.03.03.

[81] Z. Mehmood, T. Mahmood, and M. A. Javid, “Content-based image retrieval and semantic automatic image annotation based on the weighted average of triangular histograms using support vector machine,” *Applied Intelligence*, vol. 48, no. 1, pp. 166–181, 2017, doi: 10.1007/s10489-017-0957-5.

[82] A. Ali and S. Sharma, “Content based image retrieval using feature extraction with machine learning,” *International Conference on Intelligent Computing and Control Systems, ICICCS 2017*, vol. 2018-Janua, pp. 1048–1053, 2017, doi: 10.1109/ICCONS.2017.8250625.

[83] A. Olaode and G. Naghdy, “Review of the application of machine learning to the automatic semantic annotation of images,” *IET Image Process*, vol. 13, no. 8, pp. 1232–1245, 2019, doi: 10.1049/iet-ipr.2018.6153.

[84] R. Bibi, Z. Mehmood, A. Munshi, R. M. Yousaf, and S. S. Ahmed, “Deep features optimization based on a transfer learning, genetic algorithm, and extreme learning machine for robust content-based image retrieval,” *PLoS One*, vol. 17, no. 10, p. e0274764, Oct. 2022, doi: 10.1371/JOURNAL.PONE.0274764.

[85] M. Fathian, F. Akhlaghian Tab, K. Moradi, and S. Saien, “A learning automata framework based on relevance feedback for content-based image retrieval,” *International Journal of Machine Learning and Cybernetics*, vol. 9, no. 9, pp. 1457–1472, 2017, doi: 10.1007/s13042-017-0656-x.

[86] S. Jabeen, Z. Mehmood, T. Mahmood, T. Saba, A. Rehman, and M. T. Mahmood, “An effective content-based image retrieval technique for image visuals representation based on the bag-of-visual-words model,” *PLoS One*, vol. 13, no. 4, pp. 1–24, 2018, doi: 10.1371/journal.pone.0194526.

[87] A. Qayyum, S. M. Anwar, M. Awais, and M. Majid, “Medical image retrieval using deep convolutional neural network,” *Neurocomputing*, vol. 266, pp. 8–20, 2017, doi: 10.1016/j.neucom.2017.05.025.

[88] C. Bai, L. Huang, X. Pan, J. Zheng, and S. Chen, “Optimization of deep convolutional neural network for large scale image retrieval,” *Neurocomputing*, vol. 303, pp. 60–67, 2018, doi: 10.1016/j.neucom.2018.04.034.

[89] S. Hamreras, R. Benıtez-Rochel, B. Boucheham, M. A. Molina-Cabello, and E. Lopez-Rubio, “Content Based Image Retrieval by Convolutional Neural Networks,” *Int Work Conf Interp Nat Artif Comput*, vol. 2, pp. 277–286, 2019, doi: 10.1007/978-3-030-19651-6.

[90] J. R. Smith and C. S. Li, “Decoding image semantics using composite region templates,” *Proceedings - IEEE Workshop on Content-Based Access of Image and Video Libraries, CBAIVL 1998*, pp. 9–13, 1998, doi: 10.1109/IVL.1998.694467.
[91] Y. Zhuang, X. Liu, and Y. Pan, “Apply Semantic Template to Support Content-based Image Retrieval,” *Storage and Retrieval for Media Databases 2000*, vol. 3972, pp. 442–449, 1999, doi: 10.1117/12.373576.
[92] S. F. Chang, W. Chen, and H. Sundaram, “Semantic visual templates: linking visual features to semantics,” *IEEE International Conference on Image Processing*, vol. 3, pp. 531–535, 1998, doi: 10.1109/icip.1998.727321.
[93] Y. Liu, Y. Huang, S. Zhang, D. Zhang, and N. Ling, “Integrating object ontology and region semantic template for crime scene investigation image retrieval,” *12th IEEE Conference on Industrial Electronics and Applications, ICIEA 2017*, vol. 2018-Febru, pp. 149–153, 2017, doi: 10.1109/ICIEA.2017.8282831.
[94] Y. L. Qi, G. S. Zhang, and Y. L. Li, “A new method to generate semantic templates based on multilayer perceptron,” *Lecture Notes in Electrical Engineering*, vol. 382, pp. 35–41, 2016, doi: 10.1007/978-981-10-0740-8_5.
[95] K. Chu and G. H. Liu, “Image Retrieval Based on a Multi-Integration Features Model,” *Math Probl Eng*, vol. 2020, 2020, doi: 10.1155/2020/1461459.
[96] M. Ghahremani, H. Ghadiri, and M. Hamghalam, “Local features integration for content-based image retrieval based on color, texture, and shape,” *Multimed Tools Appl*, vol. 80, no. 18, pp. 28245–28263, Jul. 2021, doi: 10.1007/s11042-021-10895-z.
[97] E. R. Vimina and M. O. Divya, “Maximal multi-channel local binary pattern with colour information for CBIR,” *Multimed Tools Appl*, 2020, doi: 10.1007/s11042-020-09207-8.
[98] D. Latha and C. J. J. Sheela, “Enhanced hybrid CBIR based on multichannel LBP oriented color descriptor and HSV color statistical feature,” *Multimed Tools Appl*, vol. 81, no. 17, pp. 23801–23818, Jul. 2022, doi: 10.1007/S11042-022-12568-X/FIGURES/12.
[99] N. Ali *et al.*, “A novel image retrieval based on visual words integration of SIFT and SURF,” *PLoS One*, vol. 11, no. 6, pp. 1–20, 2016, doi: 10.1371/journal.pone.0157428.
[100] Y. D. Kenchappa and K. Kwadiki, “Content-based image retrieval using integrated features and multi-subspace randomization and collaboration,” *International Journal of System Assurance Engineering and Management*, vol. 13, no. 5, pp. 2540–2550, Oct. 2022, doi: 10.1007/S13198-022-01663-9/TABLES/3.
[101] K. T. Ahmed, S. Ummesafi, and A. Iqbal, “Content based image retrieval using image features information fusion,” *Information Fusion*, vol. 51, pp. 76–99, 2019, doi: 10.1016/j.inffus.2018.11.004.
[102] D. Niu, X. Zhao, X. Lin, and C. Zhang, “A novel image retrieval method based on multi-features fusion,” *Signal Process Image Commun*, vol. 87, Sep. 2020, doi: 10.1016/j.image.2020.115911.
[103] S. Sikandar, R. Mahum, and A. M. Alsalman, “A Novel Hybrid Approach for a Content-Based Image Retrieval Using Feature Fusion,” *Applied Sciences 2023, Vol. 13, Page 4581*, vol. 13, no. 7, p. 4581, Apr. 2023, doi: 10.3390/APP13074581.
[104] L. Yu, X. Xia, K. Zhou, and L. Zhao, “Affine invariant fusion feature extraction based on geometry descriptor and BIT for object recognition,” *IET Image Process*, vol. 13, no. 1, pp. 57–72, 2019, doi: 10.1049/iet-ipr.2018.5488.
[105] D. Kishore and C. S. Rao, “Content-Based Image Retrieval System Based on Fusion of Wavelet Transform, Texture and Shape Features,” *Mathematical Modelling of*

Engineering Problems, vol. 8, no. 1, pp. 110–116, 2021, doi: 10.18280/MMEP.080114.

[106] B. H. Yuan and G. H. Liu, "Image retrieval based on gradient-structures histogram," *Neural Comput Appl*, vol. 32, no. 15, pp. 11717–11727, 2020, doi: 10.1007/s00521-019-04657-0.

[107] S. Sathiamoorthy and M. Natarajan, "An efficient content based image retrieval using enhanced multi-trend structure descriptor," *SN Appl Sci*, vol. 2, no. 2, 2020, doi: 10.1007/s42452-020-1941-y.

[108] M. Natarajan and S. Sathiamoorthy, "Wavelet Based Multi-Trend Structure Descriptor for Effective Image Retrieval," *Proceedings of the 4th International Conference on Communication and Electronics Systems, ICCES 2019*, no. Icces, pp. 2116–2122, 2019, doi: 10.1109/ICCES45898.2019.9002388.

[109] X. Wang and Z. Wang, "A novel method for image retrieval based on structure elements' descriptor," *J Vis Commun Image Represent*, vol. 24, no. 1, pp. 63–74, 2013, doi: 10.1016/j.jvcir.2012.10.003.

[110] Z. Zeng, "A novel local structure descriptor for color image retrieval," *Information*, vol. 7, no. 9, 2016, doi: 10.3390/info7010009.

[111] W. Song *et al.*, "Taking advantage of multi-regions-based diagonal texture structure descriptor for image retrieval," *Expert Syst Appl*, vol. 96, pp. 347–357, Apr. 2018, doi: 10.1016/j.eswa.2017.12.006.

[112] T. Ma, J. Ma, and K. Yu, "A local feature descriptor based on oriented structure maps with guided filtering for multispectral remote sensing image matching," *Remote Sens (Basel)*, vol. 11, no. 8, 2019, doi: 10.3390/rs11080951.

[113] H. Weng, J. Liu, and B. Luo, "Heterogeneous image retrieval based on structural information," *Signal Image Video Process*, vol. 16, no. 4, pp. 1117–1125, Jun. 2022, doi: 10.1007/S11760-021-02061-7/FIGURES/4.

[114] H. Chugh *et al.*, "An Image Retrieval Framework Design Analysis Using Saliency Structure and Color Difference Histogram," *Sustainability 2022, Vol. 14, Page 10357*, vol. 14, no. 16, p. 10357, Aug. 2022, doi: 10.3390/SU141610357.

[115] B. Julesz, "A theory of preattentive texture discrimination based on first-order statistics of textons.," *Biological Cybern*, vol. 41, no. 2, pp. 131–138, 1981.

[116] B. Khaldi, O. Aiadi, and K. M. Lamine, "Image representation using complete multi-texton histogram," *Multimed Tools Appl*, vol. 79, no. 11–12, pp. 8267–8285, 2020, doi: 10.1007/s11042-019-08350-1.

[117] A. Raza, T. Nawaz, H. Dawood, and H. Dawood, "Square texton histogram features for image retrieval," *Multimed Tools Appl*, vol. 78, no. 3, pp. 2719–2746, 2019, doi: 10.1007/s11042-018-5795-x.

[118] A. Raza, H. Dawood, H. Dawood, S. Shabbir, R. Mehboob, and A. Banjar, "Correlated primary visual texton histogram features for content base image retrieval," *IEEE Access*, vol. 6, pp. 46595–46616, 2018, doi: 10.1109/ACCESS.2018.2866091.

[119] A. Bala and T. Kaur, "Local texton XOR patterns: A new feature descriptor for content-based image retrieval," *Engineering Science and Technology, an International Journal*, vol. 19, no. 1, pp. 101–112, 2016, doi: 10.1016/j.jestch.2015.06.008.

[120] M. Vijayashanthi, "MAGNITUDE-BASED TWIN TEXTON CO-OCCURRENCE MATRIX FOR IMAGE RETRIEVAL," *J Theor Appl Inf Technol*, vol. 15, no. 11, 2022, Accessed: Apr. 25, 2023. [Online]. Available: www.jatit.org

[121] G. H. Liu, Z. Y. Li, L. Zhang, and Y. Xu, "Image retrieval based on micro-structure descriptor," *Pattern Recognit*, vol. 44, no. 9, pp. 2123–2133, 2011, doi: 10.1016/j.patcog.2011.02.003.

[122] H. Dawood, M. H. Alkinani, A. Raza, H. Dawood, R. Mehboob, and S. Shabbir, “Correlated microstructure descriptor for image retrieval,” *IEEE Access*, vol. 7, pp. 55206–55228, 2019, doi: 10.1109/ACCESS.2019.2911954.
[123] S. Umamaheswaran, R. Lakshmanan, V. Vinothkumar, K. S. Arvind, and S. Nagarajan, “New and robust composite micro structure descriptor (CMSD) for CBIR,” *Int J Speech Technol*, vol. 23, no. 2, pp. 243–249, 2019, doi: 10.1007/s10772-019-09663-0.
[124] G. H. Liu and J. Y. Yang, “Deep-seated features histogram: A novel image retrieval method,” *Pattern Recognit*, vol. 116, Aug. 2021, doi: 10.1016/j.patcog.2021.107926.
[125] G. H. Liu and J. Y. Yang, “Deep-seated features histogram: A novel image retrieval method,” *Pattern Recognit*, vol. 116, Aug. 2021, doi: 10.1016/j.patcog.2021.107926.
[126] E. Collins and S. Susstrunk, “Deep Feature Factorization for Content-Based Image Retrieval and Localization,” *Proceedings - International Conference on Image Processing, ICIP*, vol. 2019-September, pp. 874–878, Sep. 2019, doi: 10.1109/ICIP.2019.8802980.
[127] G. H. Liu and J. Y. Yang, “Exploiting deep textures for image retrieval,” *International Journal of Machine Learning and Cybernetics*, vol. 14, no. 2, pp. 483–494, Feb. 2023, doi: 10.1007/S13042-022-01645-0/FIGURES/7.
[128] W. Wei, W. Wang, Y. Yang, and Y. Wang, “A novel color image retrieval method based on texture and deep features,” *Multimed Tools Appl*, vol. 81, no. 1, pp. 659–679, Jan. 2022, doi: 10.1007/S11042-021-11198-Z/FIGURES/11.
[129] S. Hussain, M. A. Zia, and W. Arshad, “Additive deep feature optimization for semantic image retrieval,” *Expert Syst Appl*, vol. 170, p. 114545, May 2021, doi: 10.1016/J.ESWA.2020.114545.
[130] S. Kumar, M. K. Singh, and M. Mishra, “Efficient Deep Feature Based Semantic Image Retrieval,” *Neural Process Lett*, pp. 1–24, Jan. 2023, doi: 10.1007/S11063-022-11079-Y/FIGURES/19.
[131] A. S. Tarawneh, C. Celik, A. B. Hassanat, and D. Chetverikov, “Detailed investigation of deep features with sparse representation and dimensionality reduction in CBIR: A comparative study,” *Intelligent Data Analysis*, vol. 24, no. 1, pp. 47–68, Jan. 2020, doi: 10.3233/IDA-184411.
[132] A. Alzu’bi, A. Amira, and N. Ramzan, “Content-based image retrieval with compact deep convolutional features,” *Neurocomputing*, vol. 249, pp. 95–105, 2017, doi: 10.1016/j.neucom.2017.03.072.
[133] F. Huang, C. Jin, Y. Zhang, K. Weng, T. Zhang, and W. Fan, “Sketch-based image retrieval with deep visual semantic descriptor,” *Pattern Recognit*, vol. 76, pp. 537–548, 2018, doi: 10.1016/j.patcog.2017.11.032.
[134] K. Mahantesh and A. Shubha Rao, “Content Based Image Retrieval - Inspired by Computer Vision Deep Learning Techniques,” *4th International Conference on Electrical, Electronics, Communication, Computer Technologies and Optimization Techniques, ICEECCOT 2019*, pp. 371–377, 2019, doi: 10.1109/ICEECCOT46775.2019.9114610.
[135] Y. Deldjoo, M. Elahi, M. Quadrana, and P. Cremonesi, “Using visual features based on MPEG-7 and deep learning for movie recommendation,” *Int J Multimed Inf Retr*, vol. 7, no. 4, pp. 207–219, 2018, doi: 10.1007/s13735-018-0155-1.
[136] S. Jardim, J. António, C. Mora, and A. Almeida, “A Novel Trademark Image Retrieval System Based on Multi-Feature Extraction and Deep Networks,” *Journal of Imaging 2022, Vol. 8, Page 238*, vol. 8, no. 9, p. 238, Sep. 2022, doi: 10.3390/JIMAGING8090238.

[137] N. Messina, G. Amato, F. Carrara, F. Falchi, and C. Gennaro, “Learning visual features for relational CBIR,” *Int J Multimed Inf Retr*, vol. 9, no. 2, pp. 113–124, 2019, doi: 10.1007/s13735-019-00178-7.
[138] A. B. Yandex and V. Lempitsky, “Aggregating local deep features for image retrieval,” *Proceedings of the IEEE International Conference on Computer Vision*, vol. 2015 Inter, pp. 1269–1277, 2015, doi: 10.1109/ICCV.2015.150.
[139] J. Z. Wang, J. Li, and G. Wiederhold, “SIMPLIcity: Semantics-sensitive integrated matching for picture libraries,” *IEEE Trans Pattern Anal Mach Intell*, vol. 23, no. 9, pp. 947–963, 2001.
[140] J. W. J. Z. Li, “Real-time Computerized Annotation of Pictures,” *IEEE Trans Pattern Anal Mach Intell*, vol. 30, no. 6, pp. 985–1002, 2008.
[141] Corel Dataset, “Corel-5K.” https://github.com/watersink/Corel5K
[142] Corel Dataset, “Corel-10K.” http://www.ci.gxnu.edu.cn/cbir/Dataset.aspx
[143] Caltech, “Caltech-101.” http://www.vision.caltech.edu/Image_Datasets/Caltech101/
[144] Caltech, “Caltech-256.” http://www.vision.caltech.edu/Image_Datasets/Caltech256/
[145] Alex Krizhevsky, V. Nair, and G. Hinton, “CIFAR-10.” http://www.cs.toronto.edu/~kriz/cifar.html
[146] Alex Krizhevsky, V. Nair, and G. Hinton, “CIFAR-100.” http://www.cs.toronto.edu/~kriz/cifar.html
[147] H. Jegou, M. Douze, and C. Schmid, “INRIA Holidays.” http://lear.inrialpes.fr/people/jegou/data.php
[148] PASCAL, “Pascal (VOC2012).” http://host.robots.ox.ac.uk/pascal/VOC/
[149] Stanford Vision Lab, Stanford University, and Princeton University, “ILSVRC (ImageNet).” http://www.image-net.org/about-stats
[150] GauravSingh, “texture dataset,” *https://www.kaggle.com/gauravsingh69/texture-dataset*, Oct. 31, 2019.
[151] Diego Vallejo, “texture,” *https://www.kaggle.com/dvallejin/texture*, Jan. 13, 2020.
[152] Z. G. Liu, Y. Yang, and X. H. Ji, “Flame detection algorithm based on a saliency detection technique and the uniform local binary pattern in the YCbCr color space,” *Signal Image Video Process*, vol. 10, no. 2, pp. 277–284, Feb. 2016, doi: 10.1007/s11760-014-0738-0.
[153] S. K. Mishra, K. K. Singh, R. Dixit, and M. K. Bajpai, “Design of Fractional Calculus based differentiator for edge detection in color images,” *Multimed Tools Appl*, vol. 80, no. 19, pp. 29965–29983, Aug. 2021, doi: 10.1007/s11042-021-11187-2.
[154] N. Aboutabit, “A new construction of an image edge detection mask based on Caputo–Fabrizio fractional derivative,” *Visual Computer*, vol. 37, no. 6, pp. 1545–1557, Jun. 2021, doi: 10.1007/s00371-020-01896-4.
[155] J. E. Lavín-Delgado, J. E. Solís-Pérez, J. F. Gómez-Aguilar, and R. F. Escobar-Jiménez, “A New Fractional-Order Mask for Image Edge Detection Based on Caputo–Fabrizio Fractional-Order Derivative Without Singular Kernel,” *Circuits Syst Signal Process*, vol. 39, no. 3, pp. 1419–1448, Mar. 2020, doi: 10.1007/s00034-019-01200-3.
[156] Dacheng Tao, “Corel-DB,” *https://sites.google.com/site/dctresearch/Home/content-based-image-retrieval*, Jul. 2009.

Anexos

Anexo A. Primera página del Resumen publicado en MICAI 2019

Semantic Improvements to MPEG-7 Descriptors for Content-Based Image Retrieval

K. Salvador Aguilar D. [1]

[1] Tecnológico Nacional de México/ CENIDET, Computer Science Department, Cuernavaca, Morelos, México
kevin.aguilar17ca@cenidet.edu.mx

Abstract. This abstract shows a brief description of the principal parts of research project called: "Improvements to Recommended Descriptors by the MPEG-7 standard for content-based image retrieval", who is carried out for obtain the degree of doctor, in the department of computational sciences in CENIDET.

Keywords: MPEG-7, CBIR, Semantic, Descriptors.

1 Motivation

The file multimedia storage is increasing; therefore, the need arises to efficient filter, search and identify similar visual information [2]. The Moving Picture Experts Group (MPEG) developed the MPEG-7 standard (Multimedia content description interface [2]) to solve this problem and its applications are varied, like [3]–[8]. One of the applications that are being given to the visual descriptors of this standard is in Content-Based Image Retrieval (CBIR) systems [4], [6], but the descriptors present some problems in image recovery among millions of pictures [5], [9], [10].
Attempts have been made to propose modifications to these descriptors [10], [11], however, the methods are very mathematical and still present problems in recovery: the results only are similar in one of their characteristics such as color or shape [5]. These descriptors could be improved by introducing semantics to avoid those errors that are only similar in some characteristics and do not consider the semantic content of the image.

2 Previous works in the area

Currently, several MPEG-7 standard descriptors are still being implemented for images [12], as well as for video [13] and audio [14]. Similarly, the CBIR systems are still applied to solve many problems and work has been done to improve them as shown in [15], where is proposed a novel CBIR technique based on the visual words fusion of Speeded-Up Robust Features (SURF) and Fast REtina Keypoint (FREAK) feature descriptors. Other example is [16], this work focuses on a uniform partitioning scheme

Anexo B. Primera página de Artículo para (4° JCYTA)

Jornada de Ciencia y Tecnología Aplicada
Vol. 3, Núm. 1, Enero – Junio 2020. ISSN en trámite

Evaluación Objetiva de Descriptores Visuales del Estándar MPEG-7 en la Recuperación de Imágenes

K. Salvador Aguilar Domínguez, Manuel Mejía Lavalle

Tecnológico Nacional de México / CENIDET
Cuernavaca, Morelos, México;
e-mail: {kevin.aguilar17ca, mlavalle}@cenidet.edu.mx

Resumen: En este trabajo se implementaron métricas de evaluación para los descriptores MPEG-7 presentes en LIRe y se propuso un ajuste para la métrica sugerida por el estándar, con la finalidad de obtener una evaluación objetiva de los descriptores y mejorar la evaluación realizada considerando la cantidad de imágenes recuperadas evitando así castigar la recuperación por ello. Los resultados mostraron cual descriptor presentó una mejor recuperación por si solo y la diferencia entre la evaluación con las diferentes métricas y el ajuste propuesto., la cual aparenta ser una buena alternativa de evaluación para casos donde las imágenes recuperadas sean menores a la clase de la consulta.

Palabras clave: MPEG-7, Descriptores, CBIR, Multimedia.

1. INTRODUCCIÓN

El almacenamiento de archivos multimedia está aumentando; por lo tanto, surge la necesidad de filtrar eficientemente, buscar e identificar información visual similar (J. M. Martínez, 2004). *Moving Picture Experts Group* (MPEG) desarrolló el estándar MPEG-7 (interfaz de descripción de contenido multimedia) para resolver este problema y sus aplicaciones son variadas, como(Hyun, Kim, & Oh, 2015; Mejía-Lavalle, Pérez Lara, & Ruiz Ascencio, 2013; Pattanaik & G. Bhalke, 2012; Sikora, 2001; Tyagi & Tyagi, 2017; Vertan, Badea, Florea, Florea, & Bădoiu, 2017). Una de las aplicaciones que se están dando a los descriptores visuales de este estándar es en los sistemas de recuperación de imágenes basadas en contenido (CBIR) (Hyun et al., 2015; Pattanaik & G. Bhalke, 2012), sin embargo, a pesar de que el estándar muestra información de estos descriptores, aún faltan evaluaciones objetivas de estos descriptores en la recuperación de imágenes basados en contenido con otros *Datasets* aparte de los que se presenta el estándar.

Aunque el estándar lleva más de 10 años que se presentó, actualmente, todavía se están implementando varios descriptores MPEG-7 para imágenes (Georgescu, Răducanu, & Datcu, 2017), así como para video (Duan et al., 2019) y audio (Lee & Lee, 2018). Del mismo modo, los sistemas CBIR todavía se aplican para resolver muchos problemas y se ha trabajado para mejorarlos como se muestra en (Jabeen et al., 2018), donde se propone una nueva técnica CBIR basada en la fusión de palabras visuales de Características Robustas Aceleradas (SURF) y Descriptores de funciones de REtina *Keypoint* (FREAK) rápidos. Otro ejemplo es (Fadaei, Amirfattahi, & Ahmadzadeh, 2017), este trabajo se centra en un esquema de partición uniforme que se aplica en el espacio de color *Hue*, *Saturation* and *Value* (HSV) para extraer las características del Descriptor de color dominante (DCD); este esquema CBIR, las características DCD se extrajeron inicialmente como características de color, y luego se aplicó una medida de similitud apropiada. Entre otros trabajos que resuelven otros problemas (Dash, Mukhopadhyay, & Das Gupta, 2015; Huang et al., 2018; Mohamadzadeh & Farsi, 2016; Wu, Xiao, & Hong, 2018; Xia, Zhu, Sun, Qin, & Ren, 2018).

A lo largo del tiempo se han creado diversos sistemas CBIR, que implementan descriptores del estándar. Uno de ellos es LIRe (Lux & Marques, 2013), que implementa tres descriptores del estándar. Sin embargo, aunque este sistema los implementa carece de una métrica de evaluación para realizar una comparación objetiva entre descriptores. La Sección 3 muestra más información sobre LIRe.

En el presente trabajo se implementa una evaluación objetiva de estos tres descriptores de LIRe y una propuesta de mejora para la métrica *Average Normalized Modified Retrieval Rank* (*ANMRR*) que llamamos *MEPRO*. El artículo se organiza de la siguiente manera: en la Sección 2 se explica el estándar MPEG-7; la Sección 3 describe el sistema utilizado LIRe; en la Sección 4 se presentan las métricas utilizadas, junto a la mejora propuesta; la experimentación y resultados obtenidos se presentan en la Sección 5; y finalmente las conclusiones y el trabajo futuro son mencionados en la Sección 6.

2. ESTANDAR MPEG-7

MPEG-7, un estándar desarrollado por la *International Standards Organization* (ISO) y la *International Electrotechnical Commission* (IEC). MPEG-7 describe el contenido multimedia en varios niveles, que incluyen características, estructura, semántica, modelos, colecciones y otros metadatos inmutables relacionados con la descripción

Anexo C. Primera página de Artículo para (ICMEAE 2020)

2020 International Conference on Mechatronics, Electronics and Automotive Engineering (ICMEAE)

2020 International Conference on Mechatronics, Electronics and Automotive Engineering (ICMEAE) | 978-1-7281-9904-7/20/$31.00 ©2020 IEEE | DOI: 10.1109/ICMEAE51770.2020.00017

Objective Evaluation of MPEG-7 Visual Descriptors in CBIR Systems

K. Salvador Aguilar-Domínguez[1], Manuel Mejía-Lavalle[1],
Gerardo Reyes-Salgado[1], Osslan Osiris Vergara-Villegas[2]
[1]*Tecnológico Nacional de México / CENIDET, Cuernavaca, Morelos, México*
{kevin.aguilar17ca, mlavalle, greyes }@cenidet.edu.mx
[2]*Universidad Autónoma de Ciudad Juárez, Ciudad Juárez, Chihuahua, México;*
overgara@uacj.mx

Abstract

In this work, we present an objective evaluation using metrics, for the evaluation we implemented three MPEG-7 descriptors present in LIRe and we proposed an adjustment for the metric ANMRR. Since this metric punishes the evaluation when the total recovered images are less than the class of the query image. To improve the ANMRR metric evaluation, we proposed an adjustment considering the number of recovered images. The results showed which descriptor presented a better evaluation in each one metrics, as well as the difference between the most popular evaluation metrics used in image retrieval systems and the proposed ANMRR metric. This simple adjustment of ANMRR appears to be a good evaluation alternative for cases where the recovered images are less than the class of the query.

***Keyword**: MPEG-7, Visual Descriptors, CBIR, Metrics*

1. Introduction

The storage of multimedia files is increasing; therefore, the need to efficiently filter, search and identify similar visual information arises [1], Currently there are works that address this problem in different ways, concentrating mainly on: video as in [2] that make a semantic analysis in soccer videos; or in images that are the most relevant to this work. Moving Picture Experts Group (MPEG) developed the MPEG-7 standard (multimedia content description interface), to solve this problem and its applications are varied, such as [3]–[8]. One of the applications that are being given to the visual descriptors of this standard is in content-based image retrieval systems (CBIR) [3], [5]. However, the standard shows information on these descriptors, don't present an objective evaluations in content-based image retrieval with other Datasets apart from those presented by the standard.

some current works still use MPEG-7 descriptors, in [9] used visual descriptors for multispectral earth observation image analysis, as well as, in [10] used the compact video descriptors and [11] used the audio descriptors.

CBIR systems are still applied to solve many problems and work has been done to improve them as shown in [12], where a new CBIR technique and descriptor is proposed using feature integration theory, whit a novel visual feature descriptor, namely, a multi-integration features histogram, is proposed for image representation and content-based image retrieval. Another example is [13], This work focuses on a uniform partition scheme that is applied in the Hue, Saturation and Value (HSV) color space to extract the characteristics of the Dominant Color Descriptor (DCD); the DCD features were extracted as color features, and then an appropriate similarity measure was applied. Among other jobs that solve other problems: [14]–[18]. Over time, various CBIR systems have been created, which implement descriptors of the standard. One of them is LIRe [19], implements three descriptors of the standard. However, although this system implements them, it lacks an evaluation metric to make an objective comparison between descriptors. Section 3 shows more information about LIRe.

This work implements an objective evaluation of these three LIRe descriptors and a proposal for improvement for the Average Normalized Modified Retrieval Rank (ANMRR) metric that we called Average Normalized Modified Retrieval Rank in Short Retrievals (ANMRR-SR). The article is organized as follows: Section 2 explains the MPEG-7 standard; Section 3 describes the LIRe system used; Section 4 presents the metrics used, together with the proposed improvement; the experimentation and results obtained are presented in Section 5; and finally the conclusions and future work are mentioned in Section 6.

2573-3001/20/$31.00 ©2020 IEEE
DOI 10.1109/ICMEAE51770.2020.00017

Anexo D. Primera página de Artículo para (5° JCYTA)

Jornada de Ciencia y Tecnología Aplicada
Vol. 3, Núm. 2, Julio - Diciembre 2020.

ISSN en trámite

Una Revisión Sobre Características de Alto Nivel en la Recuperación de imágenes Basadas en Contenido

K. Salvador Aguilar-Domínguez[*,1], Manuel Mejía-Lavalle[*,2] Andrea Magadán-Salazar[*,3], Gerardo Reyes-Salgado[*,4], José Ruiz-Ascencio[*,5], Osslan Osiris Vergara-Villegas[**]

[*]Tecnológico Nacional de México / CENIDET, Cuernavaca, Morelos, México; e-mail: {kevin.aguilar17ca[1], mlavalle[2], magadan[3], greyes[4], josera[5]}@cenidet.edu.mx

[**]Universidad Autónoma de Ciudad Juárez, Ciudad Juárez, Chihuahua, México; e-mail: overgara@uacj.mx

Resumen: En este trabajo se presenta la revisión sobre las características de alto nivel en los sistemas de recuperación basados en contenido. Con la finalidad de proporcionar una revisión completa y reciente, dando a conocer los logros en los diversos enfoques encontrados. La revisión se dividió en cuatro principales categorías, en las que se mencionan los trabajos considerados más relevantes en los últimos años, para finalmente presentar una discusión sobre su ventajas, desventajas y problemas actuales. La revisión muestra la demanda actual en la obtención de características de bajo nivel para mejorar los resultados en los sistemas de recuperación basados en contenido y se sugieren algunas posibles direcciones de investigación futuras.

Palabras claves: Recuperación de imágenes basada en contenido, Características de alto nivel, Revisión, Brecha semántica, Multimedia, Descriptores.

1. INTRODUCCIÓN

El desarrollo de herramientas como, los teléfonos inteligentes y diversos dispositivos que facilitan la adquisición de imágenes, al igual que las aplicaciones multimedia y redes sociales, han promovido una amplia gama de imágenes con contenido diverso. Esto provoca un gran aumento en el tamaño de la colección de imágenes digitales, tanto en los dispositivos personales como en la web. Los usuarios de diversos dominios necesitas herramientas eficientes de búsqueda, filtrado, identificación y recuperación de imágenes. Para este último, se han desarrollado los sistemas de recuperación de imágenes. Hay dos principales formas en las que se está abordando este tema: basado en texto y basado en contenido.

Los sistemas de recuperación basados en contenido "*Content-based image retrieval*" (CBIR), surgen debido a los problemas presentados en los sistemas basados en texto, como lo es el trabajo humano que se requiere para la anotación manual de las imágenes y lo complicado que puede ser la recuperación a partir de un texto, debido a la subjetividad humana. En los sistemas CBIR, también conocido como "*Query By Image Content*" (QBIC) (Assfalg, Bimbo and Pala, 2000), la recuperación se basa en el contenido de una imagen de consulta, es decir en sus características visuales de bajo nivel como: color, textura, forma y ubicaciones espaciales. En la actualidad los sistemas CBIR se utilizan en diversos campos como: Educación (Tang and Acton, 2003); medicina (Zin *et al.*, 2018); redes sociales (Ahmad *et al.*, 2016); motores de búsqueda (Deniziak and Michno, 2016); entre otras. Un ejemplo de sistemas CBIR, lo encontramos en el buscador web, *Google images*, en su opción de "buscar por imagen", donde se puede seleccionar una imagen desde un archivo o URL, para realizar una búsqueda de imágenes a partir de ésta.

Actualmente se han realizado diversos sistemas CBIR, con diferentes aplicaciones e intentando dar solución a algunos problemas que se han presentado a lo largo del tiempo, a su vez se han realizado diferentes estudios sobre ellos, algunos de ellos son: (Alzu'bi, Amira and Ramzan, 2015; Gandhani and Singhal, 2015; Shrivastava and Tyagi, 2015; Nalini and Malleswari, 2016; Mohamed, 2017; Patil and Agrawal, 2017; RajaSenbagam and Shanmugalakshmi, 2017; Zhou, Li and Tian, 2017).

Tyagi menciona en (Tyagi, 2017) algunos de los problemas de investigación sobre los sistemas CBIR, que son en los que se están centrando principalmente las recientes investigaciones los cuales son: Interacción del usuario; segmentación; reducción de dimensionalidad e indexación de características de la imagen; recuperación de imágenes basada en geo etiquetas; características de imagen de alto nivel; aprendizaje profundo; recuperación de imágenes basada en contenido para preservar la privacidad; y la recuperación de video basada en contenido. Este trabajo se concentra en la parte de características de imagen de alto nivel, que surge a partir de un problema conocido como la brecha semántica que es explicado con mayor detalle en la Sección 2 de este documento. A pesar de que se encuentra una cantidad considerable de literatura que abordan las características de alto nivel, no se ha encontrado un estudio o revisión reciente de estos trabajos. Esta revisión se dividirá en cuatro enfoques donde se mencionarán los trabajos más relevantes considerando el año de publicación, procedencia y la cantidad de citas.

Anexo E. Primera página de Artículo para (TS)

IIETA International Information and Engineering Technology Association

Traitement du Signal
Vol. 39, No. 1, February, 2022, pp. 21-29
Journal homepage: http://iieta.org/journals/ts

Image Description Using the Relation Between Color and Texture in Retrieval Task

Kevin Salvador Aguilar-Dom nguez*, Ra l Pinto-El as, Juan Gabriel Gonz lez-Serna, Andrea Magad n-Salazar

Department of Computer Science, National Center for Research and Technological Development, Cuernavaca 62490, Morelos, Mexico

Corresponding Author Email: kevin.aguilar17ca@cenidet.edu.mx

https://doi.org/10.18280/ts.390103

Received: 30 November 2021
Accepted: 6 January 2022

Keywords:
content-based image retrieval, image representation, microstructures, multi-integration features, texture descriptor, color descriptor

ABSTRACT

In the past years, significant efforts have been made for new theories and models of descriptors for Content-Based Image Retrieval systems and many effective descriptors, which use color and texture, have been established. This article presents the analysis and modifications of descriptors that use color and texture for the image retrieval task. To provide a complete detailed, and fair analysis, exposing weaknesses in descriptors and ideas to correct them. We evaluated descriptors that use color and texture, with image sets and metrics found in the literature. We compared classical descriptors that only use one low-level characteristic with descriptors that use color and texture. The analysis showed discrepancies between the model and the implementation of one of the descriptors, as well as the descriptors with the best performance, their main weaknesses, and complications when we trying to correct them. likewise, we present variants that improve the image retrieval in some cases.

1. INTRODUCTION

Content-based image retrieval (CBIR) systems are used in different areas of knowledge [1-7], however, there are still problems in this type of systems such as user interaction; segmentation; dimensionality reduction and Indexing; high-level image features; Deep learning; Privacy-preserving; and Video retrieval [8]. Some of these problems focus on reducing the semantic gap, which refers to the difference between what a user is looking for in the recovery and what the system retrieves [9, 10]. The semantic gap can be related to the descriptors, CBIR systems mainly use low-level features to represent images such as: color, shape, texture, and spatial position. However, users not only look for relationships in low-level features, but also look for relationships in high-level features, such as activities, places, objects, emotions, among others [8]. This difference between what the user is looking for and what the system retrieves, could be due to difficulties in representing high-level semantics, since low-level descriptors will hardly be directly related to high-level concepts [11].

Papers have been found in the literature proposing the use of various descriptors [12-22], some classical descriptors used are proposed by the MPEG-7 standard [23], as well as improvements based on visual theories. Some improvements use more than one low-level feature of the image, to obtain descriptors that represent features of a higher level, to relate more directly or better represent high-level features.

Many proposals use the relationship between texture and color. The descriptors are mainly based on "Textons", which is based on Julesz's Textons theory [24], as is the case of the descriptor "Micro-structure descriptor" (MSD) proposed by Liu et al. [25], where it uses the relationship of two low-level features, color and texture, making use of what they call "microstructures". The authors consider microstructures as an evolution to Textons since they use both color and texture. Subsequent years saw proposed improvements to this descriptor such as the "Correlated MicroStructure Descriptor" (CMSD) descriptor [26], which unlike MSD the CMSD identifies microstructures by establishing correlations between texture orientation, color, and intensity features. CMSD also obtains edge direction differently from MSD by adding 45 ° and 135 ° diagonal edges. Likewise, the "Structure Elements' Descriptor" (SED) presented by Wang, X. and Wang, Z. [27], is another example of descriptors using both color and texture, SED, a scaled invariant descriptor, is based on structures detected in a quantized image, using 5 different structure elements: 0 °, 90 °, 45 °, 135 °, and no direction. There are other descriptors in the literature that are based on feature integration theory, one of the most recent is the "Multi-Integration Features Histogram" (MIFH) descriptor [28], which uses color and edge features for image representation. Although different types of descriptors using more than one low-level feature have been proposed in the literature, they still need to be analyzed and improved, since there is not much information about them.

This paper presents an analysis of some of the descriptors present in the literature that use both color and texture to describe the image in CBIR systems, as well as different proposals for adjustments and modifications to two of the best evaluated descriptors, considering some of the weaknesses detected. The Section 2 shows the analysis performed on the descriptors; the Section 3 presents the improvement proposals, based on the weaknesses detected in the analysis, and the experiments; finally, the Section 4 lists the conclusions obtained.

2. DESCRIPTOR ANALYSIS

Four descriptors from the literature using both color and