Optimizing Image Retrieval: A Two-Step Content-Based Image Retrieval System Using Bag of Visual Words and Color Coherence Vectors

Muhammad Sauood[1], Muhammad Suzuri Hitam[2]*, Wan Nural Jawahir Hj Wan Yussof[3]
Faculty of Computer Science and Mathematics, Universiti Malaysia Terengganu, 21030 Kuala Nerus, Terengganu, Malaysia [1, 2, 3]
Department of Computer Sciences, Bahria University Lahore Campus, Lahore, Pakistan [1]

***Abstract*—Content-Based Image Retrieval (CBIR) systems play a crucial role in efficiently managing and retrieving images from large datasets based on visual content. This paper presents a novel bi-layer CBIR system that integrates Bag of Visual Words (BoVW) and Color Coherence Vector (CCV) methods to enhance image retrieval accuracy and performance by leveraging the strengths of both feature extraction techniques. In the first layer, the BoVW approach extracts local features and represents images as histograms of visual word occurrences, facilitating efficient initial filtering. In the second layer, CCV features are extracted from the top retrieved images to capture the spatial coherence of colour regions, providing a detailed colour signature. By combining the merits of both layers, the proposed system achieves higher retrieval precision and recall compared to the traditional single-layer approaches. Experimental results demonstrate the effectiveness of the bi-layer CBIR system in retrieving relevant images with improved accuracy, making it a valuable tool for application in image databases, digital libraries, and multimedia content management.**

***Keywords*—CBIR system; Bag of Visual Words; color coherence vectors, bi-layer CBIR; two-step CBIR feature fusion; feature extraction**

I. Introduction

As image data grows, the process to retrieve similar images from a large database becomes very challenging. The traditional approach to search images is known as text-based image retrieval (TBIR), which uses a technique where images are annotated automatically or manually, which is expensive in terms of labor, and it is very time-consuming.

In TBIR system, image search is based on metadata of the query image which leads to some subjective issues, such as variation in human perception of the image, misleading annotations that lead to inaccurate image retrieval results [1].

CBIR was introduced in early 1980s to overcome the disadvantages of traditional approach in TBIR. In CBIR approach, instead of giving text as a query, an image is feed into the system as a query and the system will look into the image database to retrieve for similar look and relevant images [2]. CBIR is the answer by the computer vision for difficulty in retrieving images from huge databases. CBIR has many applications like facial recognition systems [3], in the field of military and defence [4], retail catalogues and photograph archives [5], medical imaging [6], intellectual property [7], fight against crime [8], detection of nude and inappropriate material [9] and collection of art [10].

In CBIR system, there are two major processes. The first process is feature extraction, and the second process is feature similarity matching. In the first process, low-level features such as color, texture and shape, were extracted from both the query and image from the database using an appropriate feature extraction method. In the second process, distance measurement metrics is used to measure the similarity between the features from the query image and the respective features from the image database. Retrieved images are ranked according to the lowest distance value. The smaller the distance value, the more similar the image is according to the used features [11]. Thus, it is critical and important to choose appropriate feature extraction method(s).

Content-based image retrieval methods that rely on a single feature often produce suboptimal performance [12][13] because they capture only a limited aspect of image content. Recent literature also shows that combining multiple descriptors can address these limitations [14][15].

This paper proposes a bi-layer CBIR system that combines two complementary feature extraction techniques: Bag of Visual Words (BoVW) [16] and Color Coherence Vector (CCV) [17]. The BoVW model, inspired by text retrieval methods, has been widely adopted for its ability to handle local features and represent images as histograms of visual word occurrences.

However, BoVW primarily focuses on local descriptors and often overlooks the global spatial coherence of colour regions. To address this limitation, the CCV method is incorporated in the second layer to capture the coherence of colour regions, providing a more holistic representation of image content. The proposed system operates in two layers. In the first layer, images are indexed using the BoVW model, which involves extracting local feature descriptors, constructing a visual vocabulary through k-means clustering, and representing each image as a histogram of visual word occurrences [16]. This process allows for efficient initial filtering of images. In the second layer, the top retrieved images undergo further analysis using CCV features, which quantify the coherence of colour regions within the images [17]. Integrating these two methods enhances the system's ability to discriminate between visually similar images, thereby improving retrieval accuracy.

*Corresponding author.

The remainder of this paper is organized as follows. Section II reviews related work on CBIR, with an emphasis on advances in global, local, and hybrid descriptors. Section III describes the proposed two-step CBIR methodology in detail. Section IV provides an overview of the Corel database used for evaluation, while Section V presents and analyzes the experimental results. Finally, Section VI offers concluding remarks and outlines potential directions for future research.

II. Related Work

A content-based image retrieval (CBIR) method that leverages Zernike chromaticity distribution moments in opponent chromaticity space for color features and employs rotation and scale-invariant Contourlet transform descriptors for texture analysis was introduced in [18]. This combined feature set is both compact and robust, with image similarity assessed through a weighted distance measure. General structure of the CBIR approach is shown in Fig. 1.

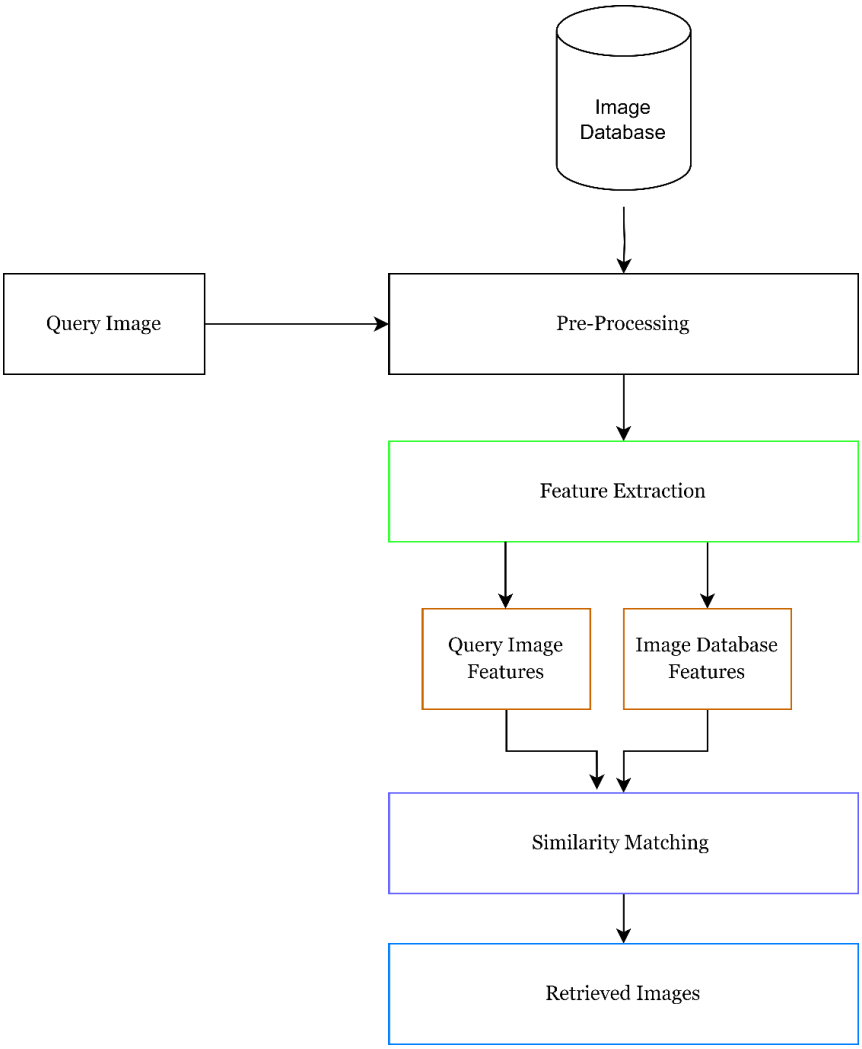

Fig. 1. General structure of the CBIR approach.

Experimental results on the COREL dataset indicate the approach achieves approximately 67% precision. The method demonstrates improved retrieval effectiveness by integrating both color and texture cues and offers resilience to variations in image brightness. However, the study provides limited insights regarding computational requirements and lacks an in-depth analysis of how parameter choices might influence retrieval outcomes. A CBIR method that fuses color histogram features with local directional pattern (LDP) descriptors to capture both color and texture characteristics of images was proposed in [19]. The approach incorporates feature normalization and a novel weighted distance metric for similarity measurement. Extensive experiments on the Wang and Corel-10000 datasets demonstrate the method's effectiveness, achieving a precision of 78.31% on the Wang dataset and a precision of 51.91% with a recall of 6.23% on the Corel-10000 dataset, outperforming several existing techniques. Strengths of this method include its robust fusion of complementary features and the clear improvement in retrieval accuracy. However, the retrieval system introduces a higher feature dimensionality, and computational efficiency is somewhat impacted compared to methods with lower-dimensional representations.

A CBIR system employing a hybrid clustering strategy that combines particle swarm optimization (PSO) with k-means clustering was presented in [20]. The method leverages color

histograms, color moments, co-occurrence matrices, and wavelet moments to extract color and texture features, and uses clustering to minimize search space and enhance retrieval speed. Experiments on the WANG dataset, which includes 1,000 images distributed across 10 classes, show the system achieves an average precision of 80.5% and a maximum class precision of 0.998 for dinosaurs, outperforming several existing methods.

Strengths of the system include its effective integration of clustering to improve retrieval accuracy and reduce computational cost at the retrieval stage. However, the system demonstrates lower precision for certain classes, such as buses and architecture, due to clustering errors, and does not incorporate shape features.

An efficient CBIR method for large databases utilizing a three-stage retrieval process based on color histograms, Gabor texture features, and Fourier descriptor-based shape features was proposed in [21]. In this approach, retrieval is performed sequentially: first selecting the top N images by color similarity, then refining to P images by texture similarity, and finally to K images using shape similarity, thereby removing the need for feature fusion and normalization. For the COREL 1k dataset, the method achieved an average precision of 76.9%. Key strengths include reduced computation time due to stage-wise narrowing of the search space and improved retrieval accuracy by integrating both global and region features.

However, the effectiveness of the system depends on careful selection of the N, P, and K parameters, and the reliability of shape features can be limited by segmentation quality, especially in complex natural images. A summary of the methods employing global features is presented in Table I.

TABLE I. COMPARISON OF CONTENT-BASED IMAGE RETRIEVAL METHODS USING GLOBAL FEATURES

Author (Year)	Key Features	Dataset	Precision	Strengths	Weaknesses
Wang et al. (2013)	Zernike chromaticity, Contourlet texture	COREL 1k	67.0%	Compact, robust, illumination-invariant	Lacks computational, parameter analysis
Zhou et al. (2018)	Color histogram + LDP	Wang, Corel-10k	78.3% (Wang), 51.9% (Corel-10k)	Strong fusion, high accuracy	Higher dimensionality, efficiency impact
Zeyad Safaa Younus et al. (2014)	PSO + k-means clustering, color & texture	WANG 1k	80.5%	Accurate, efficient clustering	Some classes are low, no shape features
Shrivastava & Tyagi (2015)	3-stage: color, texture, shape	COREL 1k	76.9%	Fast, integrates global & regional	Needs tuning, segmentation limits

A CBIR method utilizing local features by integrating scale-invariant feature transform (SIFT) and speeded up robust features (SURF) visual words within a bag-of-features framework was proposed in [22]. Separate codebooks are constructed for SIFT and SURF, and their concatenation yields a robust image representation, with similarity measured via a support vector machine (SVM) using a Hellinger kernel. Experiments on the Corel-1000, Corel-1500, and Corel-2000 datasets demonstrate mean average precision of 75.17%, 74.95%, and 65.41%, respectively, surpassing several state-of-the-art methods. The approach's main strengths are improved retrieval performance and robustness to scale, rotation, and illumination changes by combining both descriptors. However, the method increases feature dimensionality and relies on parameter choices (such as vocabulary size and clustering), whose effects on retrieval effectiveness and computational efficiency are not deeply analyzed.

A method for CBIR that combines Local Binary Pattern Variance (LBPV) and Local Intensity Order Pattern (LIOP) features was presented in [23]. By forming two compact visual dictionaries and then concatenating them, the method achieves robust representation and improved retrieval performance. Dimensionality reduction using principal component analysis (PCA) and the use of SVM for classification help manage computational load. The proposed system demonstrated high precision, 76.02% on WANG-1.5K, and was efficient in terms of computational cost. However, it is not directly applicable to multispectral images due to loss of spectral and spatial detail.

An encrypted image retrieval method for cloud computing that integrates optimized Harris corner detection with SURF features within a Bag-of-Words model and employs p-stable Locality Sensitive Hashing (LSH) for indexing was proposed in [24]. The approach extracts local feature points via an improved Harris algorithm, enhances feature description with SURF, and utilizes a chaotic encryption scheme to ensure the security of both images and indexes. Experimental evaluation on the Corel test set (1000 images, 10 categories) shows the proposed scheme achieves higher retrieval accuracy and lower search time compared to existing encrypted image retrieval methods, with a top-10 retrieval precision of up to 54.3%. Strengths of the method include increased retrieval efficiency and security through efficient local feature extraction and optimized hashing. However, retrieval precision decreases as the number of retrieved images increases, and the effectiveness of the scheme depends on parameter choices (e.g. number of hash tables, LSH functions), which are not deeply analyzed for sensitivity or scalability. Table II provides an overview of methods that utilize local features.

A CBIR system that integrates both local and global features by combining SIFT, Histogram of Oriented Gradients (HOG), and Local Binary Pattern (LBP) descriptors within a bag-of-features (BoF) framework was proposed in [25]. The approach explores both patch-based and image-based integration models, with feature codebooks constructed using a weighted K-means clustering algorithm. Experiments on the Corel-1000 dataset show that the image-based SIFT-LBP integration achieves the highest average retrieval precision of 65.7%, outperforming conventional color, texture, and shape-based approaches as well as BoF and spatial pyramid matching (SPM) baselines. Key strengths include complementary representation of image content and superior robustness against background noise. However, the system introduces a higher computational cost due to combined feature extraction, and its performance is sensitive to codebook size and feature weighting parameters, which require careful tuning.

TABLE II. COMPARISON OF CONTENT-BASED IMAGE RETRIEVAL METHODS USING LOCAL FEATURES

Author (Year)	Key Features / Approach	Dataset(s)	Precision (MAP/Top-10)	Strengths	Weaknesses
Nouman Ali et al. (2016)	SIFT + SURF (BoW), dual codebooks, SVM-Hellinger	Corel-1000/1500/2000	75.17% / 74.95% / 65.41%	High retrieval, robust to scale/rotation/lighting	High dimensionality, parameter sensitivity
Sarwar et al. (2019)	LBPV + LIOP, dual dictionaries, PCA, SVM	WANG-1.5K	76.02%	Compact, efficient, improved accuracy	Not suited for multispectral images
Jiaohua Qin et al. (2019)	Harris + SURF (BoW), p-stable LSH, chaotic encryption	Corel-1k	54.3% (Top-10)	Secure, efficient hashing, improved accuracy	Precision drops for more retrieved images, parameter choices not deeply analyzed

A multi-level colored directional motif histogram (MLCDMH) approach for CBIR that integrates both color (global) and local structural features was introduced in [26]. The method extracts directional motif patterns at multiple levels, applies a weighted neighboring similarity scheme, and fuses directional features into a single invariant feature map, generating compact histograms from all color planes. On the Corel-1000 dataset, the system achieved an average precision of 64.00% for the top 10 retrieved images and 59.60% for the top 20, surpassing several motif- and non-motif-based approaches. Strengths include robust representation of both local and global image characteristics and reduced matching overhead through single-vector fusion. However, system performance is dependent on block size and motif pattern parameters, whose sensitivity is not extensively explored.

A CBIR system that integrates Hue Saturation and Value (HSV) color histograms (global color), discrete wavelet transform (global texture), and edge histogram descriptor (local texture) features was proposed in [27]. The method forms a composite feature vector from these three descriptors and measures similarity using Manhattan distance. Evaluation on the Corel-1k dataset shows that the proposed system achieves an average precision of 73.5% for the top-20 retrieved images, outperforming several state-of-the-art CBIR methods. Key strengths include the effective fusion of color and texture features, which enhances retrieval performance, and competitive results across diverse categories. However, the approach's performance is dependent on the choice of block sizes and quantization levels, and retrieval efficiency may be affected by increased feature vector dimensionality.

A CBIR method that combines local features SURF and global features HOG using both feature fusion and visual words fusion strategies within a Bag-of-Visual-Words (BoVW) framework was proposed in [28]. The approach constructs separate vocabularies for SURF and HOG, then fuses them to form a compact and robust image representation. Experiments on the Corel-1000, Corel-1500, and Corel-5K datasets show that the visual words fusion method achieves a mean average precision of 80.61%, 76.20%, and 60.60%, respectively, outperforming state-of-the-art techniques. Strengths include significantly improved retrieval performance and reduced semantic gap through integrated representation of local and global features. However, increased vocabulary size and feature dimensionality can lead to higher computational cost, and the method's performance is sensitive to codebook size and feature selection percentage, requiring careful tuning. Table III summarizes the comparison of CBIR methods integrating local and global features.

TABLE III. COMPARISON OF CONTENT-BASED IMAGE RETRIEVAL METHODS USING HYBRID LOCAL AND GLOBAL FEATURES

Author (Year)	Key Features / Approach	Dataset(s)	Precision (MAP)	Strengths	Weaknesses
Jing Yu et al. (2013)	SIFT, HOG, LBP (BoF), patch/image-based fusion	Corel-1000	65.7% (image-based SIFT-LBP)	Complementary, robust to noise	High computation; sensitive to codebook/weights
Jitesh Pradhan et al. (2020)	MLCDMH: color & local motif fusion	Corel-1000	64.0% (top 10), 59.6% (top 20)	Robust global/local fusion; compact representation	Sensitive to block/motif size, not explored
Atif Nazir et al. (2018)	HSV hist., DWT (global), Edge hist. (local)	Corel-1k	73.5% (top 20)	Effective color/texture fusion, strong category results	Block/quantization choice, high dimensionality
Zahid Mehmood et al. (2018)	SURF (local) + HOG (global), BoVW fusion	Corel-1000/1500/5K	80.6%, 76.2%, 60.6%	High performance; reduced semantic gap	High computation; sensitive to parameters

While CBIR approaches in recent literature utilize either global features, local features or combination of both local and global, a critical analysis reveals distinct strengths and limitations in these individual methods, which directly inform the rationale for combining Bag of Visual Words (BoVW) and Color Coherence Vector (CCV) in the proposed system.

The BoVW model offers significant advantages in large-scale CBIR tasks. BoVW provides a fixed-length feature vector regardless of image dimensions or complexity, enabling efficient indexing and fast retrieval in large image databases. It leverages local descriptors that are robust to scale, rotation, and illumination changes, making it suitable for scenarios involving partial occlusion or background clutter, as local features can capture salient regions even when portions of an object are missing or obscured. Additionally, the construction of the visual vocabulary in BoVW is typically unsupervised and does not require manual annotation, as k-means clustering is applied to local feature descriptors to build the codebook. However, BoVW predominantly focuses on the distribution of local features and often neglects the broader spatial and color coherence information within the image, potentially leading to suboptimal retrieval when global color structures play a significant role in image similarity.

In contrast, the Color Coherence Vector (CCV) method excels at capturing the spatial and texture structure of images by distinguishing between coherent and incoherent pixels.

Coherent pixels are part of large, connected regions of similar colors, enabling CCV to robustly represent objects that share dominant color regions, while filtering out scattered, insignificant color areas. CCV thus complements BoVW by capturing global color and spatial coherence—information often missed by purely local approaches.

Therefore, the combination of BoVW and CCV directly addresses the limitations of each individual method. By integrating BoVW's robust local feature representation with CCV's global color and spatial coherence descriptors, the proposed CBIR system can better discriminate between images that are similar in both structure and content, even in the presence of occlusions, clutter, or variable lighting conditions. This dual-layer approach is expected to improve retrieval precision, particularly for complex or diverse image datasets, and provides a sound justification for introducing both methods in the system.

III. Proposed Methodology

In this work, we propose a two-layer Content-Based Image Retrieval (CBIR) system that integrates local and global feature analysis in a cascaded manner to improve retrieval effectiveness. The system consists of an initial retrieval layer based on the Bag of Visual Words (BoVW) model using SURF descriptors, followed by a refinement layer utilizing Color Coherence Vector (CCV) features. Both retrieval layers employ the Manhattan (City-Block) distance metric for similarity measurement to ensure consistency and interpretability throughout the retrieval process.

A. Layer 1: Initial Retrieval via Bag of Visual Words (BoVW)

Let the image database be denoted as $\mathcal{D} = \{I_1, I_2, \dots, I_N\}$ and the query image as I_Q.

Step 1: Local Feature Extraction

For each image I, local descriptors are extracted using the SURF algorithm. Let the set of descriptors for image I be $\{d_1, d_2, \dots, d_N\}$, where each $d_i \in \mathbb{R}^D$ is a D-dimensional SURF feature vector.

Step 2: Visual Vocabulary Construction

All descriptors from the entire database are pooled and clustered via k-means clustering into K clusters, where each cluster center v_k $(k = 1,2, \dots, K)$ defines a visual word. The value of K is a design parameter.

Step 3: Histogram Encoding

Each descriptor d_i from image I is assigned to the nearest cluster center (visual word) based on the Euclidean norm. The BoVW histogram for image I is defined as:

$$h_k = \sum_{i=1}^{N_I} \delta\left(\arg\min_j |d_i - v_j|^2 = k\right) \quad (1)$$

where $\delta(\cdot)$ is the Kronecker delta function which equals 1 if the condition is true and 0 otherwise.

Step 4: Similarity Computation

The query image I_Q is similarly encoded as a BoVW histogram h_Q. The similarity between h_Q and each database histogram h_I is computed using the Manhattan distance:

$$D_1(h_Q, h_I) = \sum_{k=1}^{K} |h_{Q,k} - h_{I,k}| \quad (2)$$

where $h_{Q,k}$ and $h_{I,k}$ are the k^{th} bins of the query and candidate histograms, respectively.

Step 5: Candidate Selection

The images in the database are ranked in ascending order of $D_1(h_Q, h_I)$, and the top N_{top} candidates with the smallest distances (where $N_{top} = 50$ in our implementation) are selected for the second retrieval layer.

B. Layer 2: Refinement via Color Coherence Vector (CCV)

Step 1: Color Quantization

For each candidate image (from Layer 1) and the query image, color quantization is performed to map all pixels into L discrete color levels. Let the color level of pixel i be $c_i \in \{1,2, \dots, L\}$.

Connected component analysis is performed on each quantized image to identify spatially contiguous regions of pixels with the same color level. For each color level λ, pixels are partitioned into two categories based on the size of the connected region (component) they belong to:

- Coherent pixels: Pixels that belong to a connected component (region) whose size is greater than or equal to a specified threshold T.
- Non-coherent pixels: Pixels that belong to a connected component whose size is less than T.
- Formally, for each color level λ, the numbers of coherent and non-coherent pixels are computed as:

$$C_\lambda = \sum_{i \in coh} \delta(c_i = \lambda), N_\lambda = \sum_{i \in non-coh} \delta(c_i = \lambda) \quad (3)$$

where, C_λ and N_λ are the counts of coherent and non-coherent pixels at color level λ, respectively.

Step 2: Feature Vector Formation

The CCV feature vector $v = [C_1, N_1, C_2, N_2, \dots, C_L, N_L]$ of length $2L$ is constructed for each image.

Step 3: Similarity Computation

The similarity between the CCV feature vector of the query image v_Q and each candidate image v_I is computed using the Manhattan distance:

$$D_2(v_Q, v_I) = \sum_{j=1}^{2L} |v_{Q,j} - v_{I,j}| \quad (4)$$

where, $v_{Q,j}$ and $v_{I,j}$ are the j^{th} elements of the query and candidate CCV vectors, respectively.

Step 4: Final Ranking

The candidate images are ranked based on their $D_2(v_Q, v_I)$ values in ascending order. The retrieval output is the set of top 10 images most similar to the query in terms of color coherence.

The workflow of the proposed approach is shown in Fig. 2.

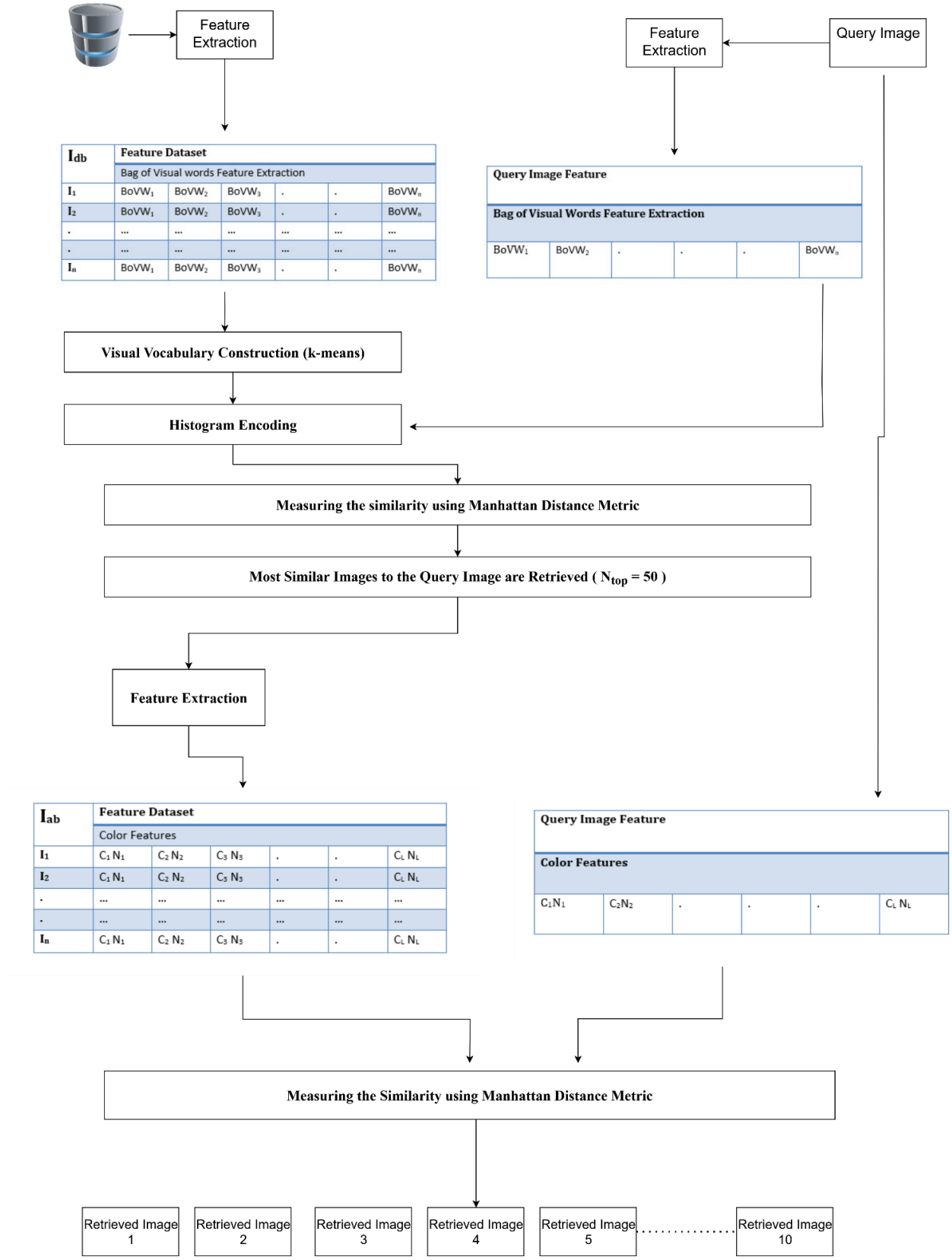

Fig. 2. Diagram of the proposed two step CBIR system.

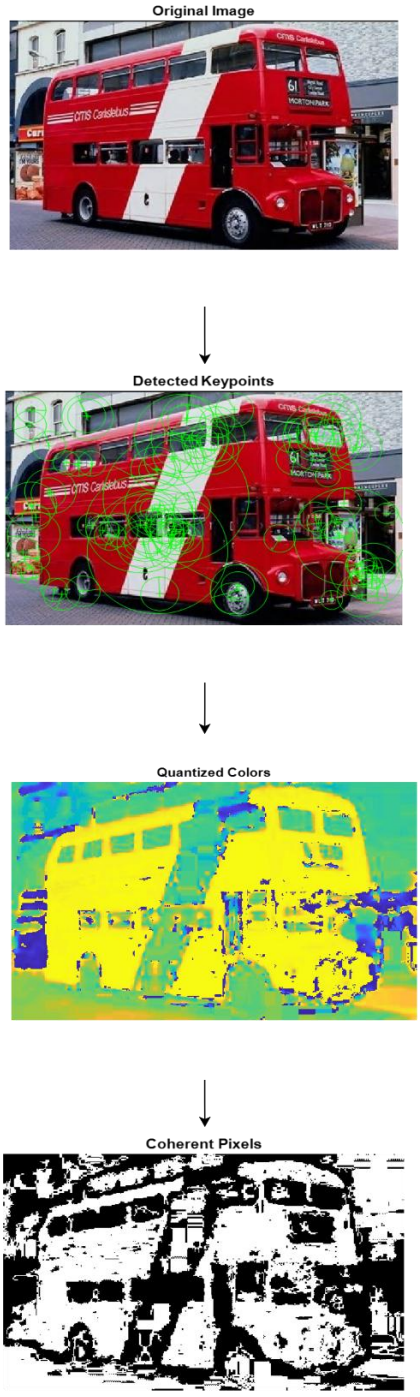

Fig. 3. Intermediate outputs of the proposed two-layer CBIR system.

Fig. 3 presents intermediate outputs from the proposed bi-layer CBIR framework. The first view shows the query image; the second depicts SURF keypoints used in the Bag-of-Visual-Words layer, which retrieves the Top-50 candidate images. The next view illustrates the Color Coherence Vector (CCV) computed in the second layer on these shortlisted images, followed by its binary coherence mask. Together, these steps show how the framework first filters candidates using local descriptors and then refines similarity using global color information.

IV. Database

The experiments are carried out on a publicly available database known as Corel-1k, which consists of 1,000 images organized into 10 categories. Every category consists of 100 images. The categories include natural scenes, everyday objects, animals, drinks, and aviation, among others. The database was selected because of its diversity and comprehensiveness. Corel-1k is a widely used benchmark image database in CBIR research, serving as a standard for evaluating retrieval accuracy and robustness. Each category in the database corresponds to different semantic concepts, facilitating the assessment of CBIR system performance. Some of the images are shown in Fig. 4 from the image database [29].

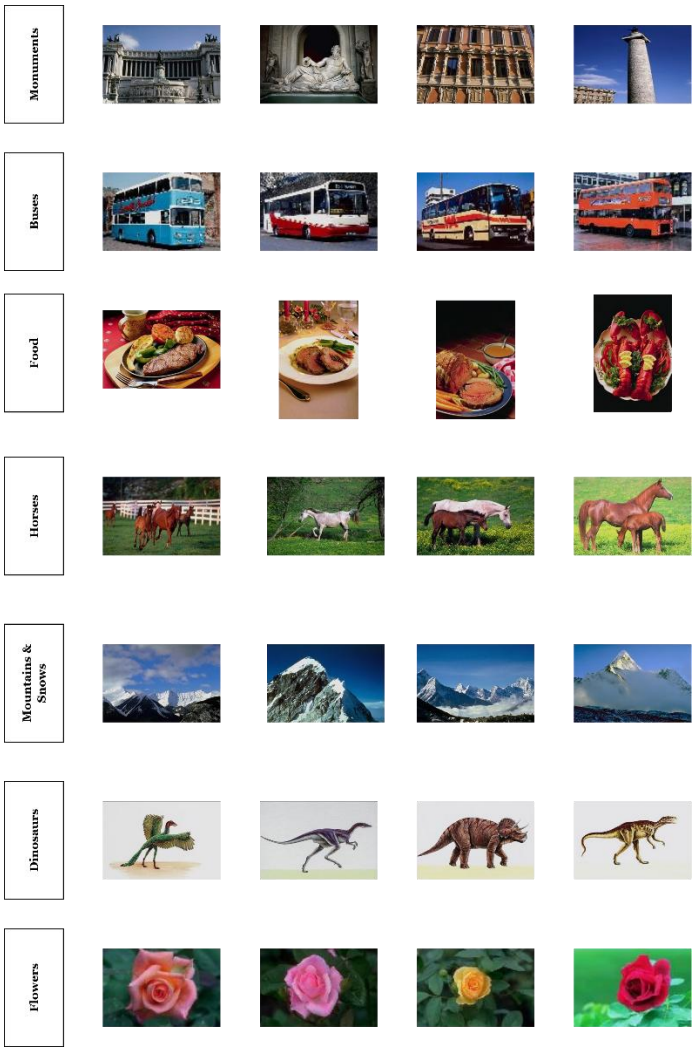

Fig. 4. Sample of images from Corel-1k dataset.

V. Experimental Results

This section evaluates the performance of the proposed novel two-layered CBIR approach. The system was assessed using the Corel-1K dataset, a benchmark frequently employed for testing CBIR methodologies due to its diverse content across

various categories. The performance metrics selected for evaluation were precision and recall, as these measures effectively indicate the system's accuracy and retrieval effectiveness.

Precision and recall are computed using standard formulas as follows:

Precision = (Number of relevant images retrieved) / (Total images retrieved).

Recall = (Number of relevant images retrieved) / (Total relevant images in the dataset).

Table IV presents the category-wise Precision@10 and Recall@10 (%) for the proposed two-layer CBIR system. The method attains an average precision of 82.2 % and a recall of 8.2 %, demonstrating that retrieved images are largely relevant even though recall is inherently limited by the top-10 evaluation protocol.

Strengths. Categories such as Buses, Dinosaurs, and Flowers achieve perfect precision, confirming that the approach effectively handles classes with uniform appearance and well-defined structure. Good results are also obtained for Horses and Elephants (P@10 = 90 %), highlighting the method's ability to manage moderate intra-class variation.

Weaknesses. Retrieval is less accurate for *Beaches* (P@10 = 45 %) and, to a lesser extent, Monuments *and* Mountains and Snow. These scene-type categories exhibit large variability in layout and background, leading to feature overlap with other classes.

Overall, the findings show that the proposed system excels in high-precision retrieval for object centric images, while scene categories remain challenging a direction for subsequent improvement.

Precision values demonstrate that most images retrieved at the top ranks are relevant. The comparatively small recall values arise from the evaluation strategy: since only the top 10 images are considered and each Corel-1K class contains 100 relevant images, even a perfect system can obtain at most 10 % recall.

The retrieval effectiveness of various CBIR methods was assessed primarily using the P@10 precision metric. The proposed approach achieved a precision rate of 82.2%, surpassing all benchmarked state-of-the-art methods considered in this comparative analysis as shown in Table V.

TABLE IV. PRECISION AND RECALL RATES FOR THE PROPOSED TWO-LAYERED CBIR SYSTEM

Categories	Precision @ 10	Recall @ 10
Buses	100	10
Dinosaurs	100	10
Horses	90	9
Elephants	90	9
Flower	100	10
Monuments	60	6
Mountains and Snow	70	7
Beaches	45	4.5
Africa	84	8.4
Food	83	8.3
Average	82.2	8.2

TABLE V. COMPARISON OF CBIR SYSTEMS IN TERMS OF PRECISION, DATASET, AND FEATURE EXTRACTION METHODS

Author(s) & Year	Features Used	Technique/Model	Distance Metric	Precision Rate
Lin et al. (2009)	CCM (color), DBPSP (texture), CHKM (color distribution)	Sequential Forward Selection (SFS)	Euclidean	72.70%
Huang et al. (2010)	Gabor texture, HSV color moment	Feature fusion	Euclidean	63.6%
Singha et al. (2012)	Color histogram, Wavelet transform	WBCHIR	Histogram Intersection	76.2%
Yu et al. (2013)	SIFT+LBP, HOG+LBP	Patch-based & image-based BoF integration	Weighted K-means	65%
Aiswarya et al. (2020)	Deep features via stacked autoencoder	Query image space generation + autoencoder	Not specified	67%
XIE et al. (2020)	Dominant color descriptor, Hu moments	Texton template-based zone detection	Not specified	80.40%
Nazir et al. (2018)	Edge histogram (local), DWT, color histogram (global)	Feature concatenation	Manhattan	73.5%
Sadique et al. (2019)	SURF, color moments (local), modified GLCM (global)	Feature concatenation	Approx. Nearest Neighbor	70.48%
Pradhan et al. (2019)	MLCDMH (local structural)	Multi-level histogram descriptor	Not specified	64%
Proposed Approach	Bag of Visual Words, Color Coherence Vector	Two-Step	Manhattan	82.2%

Among competing techniques, very few managed precision scores exceeding the 75% threshold, with the nearest competitor reaching 80.40%, and other notable methods clustered around 75–77%. Despite being close in performance, these methods ultimately fell short compared to the proposed system, highlighting the significant benefit of integrating local and global feature descriptors within a coherent retrieval framework.

Several other evaluated approaches exhibited moderate precision rates, generally between 65% and 74%, which encompassed both traditional feature engineering methods and hybrid techniques. Despite utilizing varied feature descriptors, these mid-tier methods demonstrated limitations either in the comprehensiveness of feature representation or in the consistency of similarity measurement, preventing them from achieving higher retrieval accuracy.

Methods at the lower end of performance attained precision scores below 65%, with the lowest being precisely 65%. This clearly indicates inadequate discriminative capability, likely attributable to limited descriptor diversity or suboptimal distance metrics. Although these methods frequently incorporated spatial and color information, their relatively weak retrieval results underline the necessity of employing more robust and complementary feature integration strategies.

The evident superiority of the proposed method, in terms of both absolute precision and the comparative performance margin, underscores the effectiveness of its two-phase retrieval strategy. This strategy, characterized by combining Bag of Visual Words (BoVW) and Color Coherence Vector (CCV), produces a synergistic enhancement of retrieval performance across diverse queries. The consistently superior performance relative to other benchmark methods confirms the robustness and practical applicability of the proposed CBIR approach in real-world scenarios.

To provide deeper insight into the retrieval capability and limitations of the proposed CBIR method, representative examples from two distinct categories “Dinasaurs” (high-performing) and 'Beach' (low-performing)—are presented. The precision-recall curves illustrated in Fig. 5 and Fig. 6 clearly highlight the method's superior performance for structured, visually coherent object categories, exemplified by the near-perfect retrieval results for the “Dinasaurs” class.

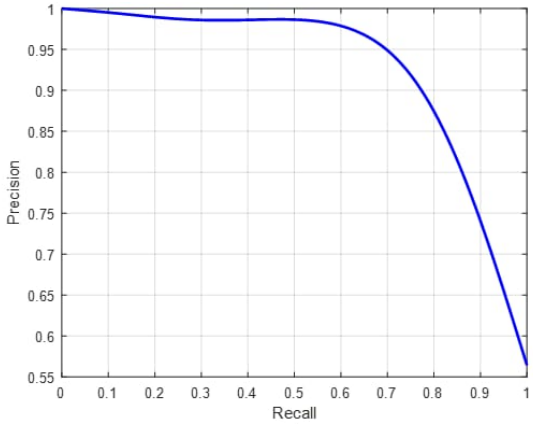

Fig. 5. Precision-recall graph for Dinasaurs class.

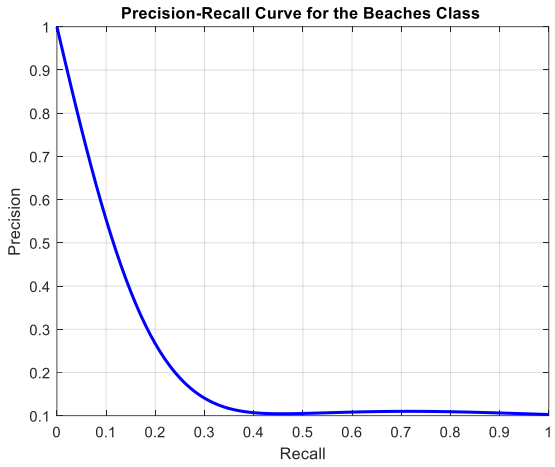

Fig. 6. Precision-recall graph for Beaches class.

Conversely, the “Beach” category exemplifies challenges inherent to semantically diverse and visually heterogeneous scenes, thus offering transparency into areas for further methodological refinement.

VI. Conclusion

This study aimed to advance the performance of CBIR systems by introducing a two-stage retrieval approach that effectively integrates local and global visual descriptors. Experimental evaluations demonstrated that the proposed framework performs competitively relative to existing state-of-the-art techniques, exhibiting improved accuracy in retrieving relevant images across diverse datasets.

By employing the Bag of Visual Words (BoVW) and Color Coherence Vector (CCV) descriptors, the proposed system successfully captures both structural and color-based visual information, significantly enhancing retrieval precision. The dual-layer architecture facilitates refined filtering and contributes positively to overall retrieval effectiveness without compromising generalization capabilities.

Despite these advantages, the method faces challenges regarding computational costs, particularly for high-resolution images and large-scale databases, as feature extraction and similarity matching become resource intensive. Future work will focus on designing lightweight variants of the framework, adopting more efficient feature-learning strategies and implementing real-time retrieval mechanisms to support interactive applications. Additional research may also explore relevance-feedback techniques and hybrid descriptors to further improve recall.

In conclusion, the developed CBIR framework offers a promising solution for practical applications, particularly in fields such as medical imaging, multimedia databases, and digital library management. Future optimizations may further enhance scalability and real-time responsiveness, positioning this methodology as a robust solution for evolving image retrieval challenges.

REFERENCES

[1] Rui, Yong, Thomas S. Huang, and Shih-Fu Chang. "Image retrieval: Current techniques, promising directions, and open issues." Journal of visual communication and image representation 10, no. 1 (1999): 39-62. https://doi.org/10.1006/jvci.1999.0413.

[2] Liu, Ying, Dengsheng Zhang, Guojun Lu, and Wei-Ying Ma. "A survey of content-based image retrieval with high-level semantics." Pattern recognition 40, no. 1 (2007): 262-282. https://doi.org/10.1016/j.patcog.2006.04.045

[3] Li Qinjun, Cui Tianwei, Zhao Yan, Wu Yuying, "Facial Recognition Technology: A Comprehensive Overview," Academic Journal of Computing & Information Science, Vol. 6, Issue 7, pp. 15-26, 2023, Francis Academic Press, UK. DOI: 10.25236/AJCIS.2023.060703.

[4] Kakizaki, K., Fukuchi, K., & Sakuma, J. (2023). Certified Defense for Content Based Image Retrieval. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) (pp. 4561-4570). IEEE.

[5] Lew, M. S., Sebe, N., Djeraba, C., & Jain, R. (2006). Content-based multimedia information retrieval: State of the art and challenges. ACM Transactions on Multimedia Computing, Communications, and Applications, 2(1), 1-19.

[6] Sudhish, D. K., Nair, L. R., & Shailesh, S. (2024). Content-based image retrieval for medical diagnosis using fuzzy clustering and deep learning. Biomedical Signal Processing and Control, 88, 105620. https://doi.org/10.1016/j.bspc.2023.105620

[7] António, J., Valente, J., Mora, C., Almeida, A., & Jardim, S. (2024). DarwinGSE: Towards better image retrieval systems for intellectual property datasets. PLOS ONE, 19(7), e0304915. https://doi.org/10.1371/journal.pone.0304915

[8] Liu, Y., Zhou, A., Xue, J., & Xu, Z. (2024). A Survey of Crime Scene Investigation Image Retrieval Using Deep Learning. Journal of Beijing Institute of Technology (English Edition), 33(4), 271-286. DOI: 10.15918/j.jbit1004-0579.2023.152

[9] Li, Xinfeng; Yang, Yuchen; Deng, Jiangyi; Yan, Chen; Chen, Yanjiao; Ji, Xiaoyu; Xu, Wenyuan. SafeGen: Mitigating Sexually Explicit Content Generation in Text-to-Image Models. In: Proceedings of ACM SIGSAC Conference on Computer and Communications Security (CCS '24), Salt Lake City, UT, USA, October 14-18, 2024. DOI: 10.1145/3658644.3670295.

[10] Castellano, G., Lella, E., & Vessio, G. (2021). Visual link retrieval and knowledge discovery in painting datasets. Multimedia Tools and Applications, 80, 6599–6616. https://doi.org/10.1007/s11042-020-09995-z

[11] Atlam, Hany Fathy, Gamal Attiya, and Nawal El-Fishawy. "Comparative study on CBIR based on color feature." International Journal of Computer Applications 78, no. 16 (2013).

[12] Al-Jubouri, Hanan, and Hongbo Du. "A Content-Based Image Retrieval Method By Exploiting Cluster Shapes." Iraqi Journal for Electrical & Electronic Engineering 14, no. 2 (2018).

[13] Sadique, Md Farhan, Bishajit Kumar Biswas, and SM Rafizul Haque. "Unsupervised content-based image retrieval technique using global and local features." In 2019 1st International Conference on Advances in Science, Engineering and Robotics Technology (ICASERT), pp. 1-6. IEEE, 2019.

[14] Ashraf, Rehan, Mudassar Ahmed, Usman Ahmad, Muhammad Asif Habib, Sohail Jabbar, and Kashif Naseer. "MDCBIR-MF: multimedia data for content-based image retrieval by using multiple features." Multimedia tools and applications 79, no. 13 (2020): 8553-8579. https://doi.org/10.1007/s11042-018-5961-1

[15] Xie, Guangyi, Baolong Guo, Zhe Huang, Yan Zheng, and Yunyi Yan. "Combination of dominant color descriptor and Hu moments in consistent zone for content based image retrieval." IEEE Access 8 (2020): 146284-146299 10.1109/ACCESS.2020.3015285.

[16] Sivic, and Zisserman. "Video Google: A text retrieval approach to object matching in videos." In Proceedings ninth IEEE international conference on computer vision, pp. 1470-1477. IEEE, 2003.

[17] Pass, Greg, Ramin Zabih, and Justin Miller. "Comparing images using color coherence vectors." In Proceedings of the fourth ACM international conference on Multimedia, pp. 65-73. 1997.

[18] Wang, Xiang-Yang, Hong-Ying Yang, and Dong-Ming Li. "A new content-based image retrieval technique using color and texture information." *Computers & Electrical Engineering*, 39.3 (2013): 746-761.

[19] Zhou, Ju-Xiang, et al. "A new fusion approach for content based image retrieval with color histogram and local directional pattern." *International Journal of Machine Learning and Cybernetics*, 9 (2018): 677-689.

[20] Younus, Z. S., Mohamad, D., Saba, T., Alkawaz, H. M., Rehman, A., Al-Rodhaan, M., & Al-Dhelaan, A. "Content-based image retrieval using PSO and k-means clustering algorithm." *Arabian Journal of Geosciences*, 8(8), 6211–6224, 2015. doi:10.1007/s12517-014-1584-7

[21] Shrivastava, N., & Tyagi, V. "An efficient technique for retrieval of color images in large databases." *Computers & Electrical Engineering*, 46, 314–327, 2015. https://doi.org/10.1016/j.compeleceng.2014.11.009

[22] Ali, N., Bajwa, K. B., Sablatnig, R., Chatzichristofis, S. A., Iqbal, Z., Rashid, M., & Habib, H. A. "A novel image retrieval based on visual words integration of SIFT and SURF." *PloS one*, 11(6), e0157428, 2016.

[23] Sarwar, A., Mehmood, Z., Saba, T., Qazi, K. A., Adnan, A., & Jamal, H. "A novel method for content-based image retrieval to improve the effectiveness of the bag-of-words model using a support vector machine." *Journal of Information Science*, 45 (1), 117–135, 2019. https://doi.org/10.1177/0165551518782825

[24] Qin, J., Li, H., Xiang, X., Tan, Y., Pan, W., Ma, W., & Xiong, N. N. "An encrypted image retrieval method based on harris corner optimization and LSH in cloud computing." *IEEE Access*, 7, 24626–24633, 2019. https://doi.org/10.1109/ACCESS.2019.2894673

[25] Yu, Jing, et al. "Feature integration analysis of bag-of-features model for image retrieval." *Neurocomputing*, 120 (2013): 355-364.

[26] Pradhan, Jitesh, et al. "Multi-level colored directional motif histograms for content-based image retrieval." *The Visual Computer*, 36.9 (2020): 1847-1868.

[27] Nazir, Atif, et al. "Content based image retrieval system by using HSV color histogram, discrete wavelet transform and edge histogram descriptor." *2018 International Conference on Computing, Mathematics and Engineering Technologies (iCoMET)*. IEEE, 2018.

[28] Mehmood, Z., Abbas, F., Mahmood, T., Javid, M. A., Rehman, A., & Nawaz, T. "Content-based image retrieval based on visual words fusion versus features fusion of local and global features." *Arabian Journal for Science and Engineering*, 43(12), 7265–7284, 2018. https://doi.org/10.1007/s13369-018-3062-0

[29] Duygulu, Pinar, Kobus Barnard, Joao FG de Freitas, and David A. Forsyth. "Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary." In Computer Vision—ECCV 2002: 7th European Conference on Computer Vision Copenhagen, Denmark, May 28–31, 2002 Proceedings, Part IV 7, pp. 97-112. Springer Berlin Heidelberg, 2002.