Journal of Information Systems Engineering and Management
2025, 10(61s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

A Classification-Based Framework for Semantic Local and Global Image Retrieval

Mohammed Salim Meflah[1], Mohammed Lamine Kherfi[2], Sihem Kechida[3]
[1]*Department of Computer Science, Faculty of Technology, University Badji Mokhtar–Annaba, Algeria*
[2]*Department of Computer Science, Faculty of Technology, University of Kasdi Merbah, Ouargla, Algeria.*
[3]*Laboratoire d'Automatique et Informatique de Guelma (LAIG), University 8 mai 1945, Guelma, Algeria*.

ARTICLE INFO

Received: 24 Dec 2024

Revised: 12 Feb 2025

Accepted: 26 Feb 202

ABSTRACT

This paper presents a novel framework for enhancing both local and global image retrieval by leveraging semantic region classification and deep learning. The core of our approach involves the unsupervised clustering of image regions into semantically meaningful categories. Each image in the database is subsequently represented by a membership weight vector indicating its affinity to these region categories, refined through a Convolutional Neural Network (CNN). This

representation enables a flexible and expressive query paradigm, allowing users to formulate precise queries by logically combining example regions using operators such as AND, OR, and NOT, as well as by specifying positive (example) and negative (counter-example) constraints. Experimental results demonstrate a substantial improvement in retrieval accuracy and user satisfaction compared to conventional methods, achieving up to a 20% increase in mean Average Precision (mAP) and confirming the effectiveness of our classification-based approach in bridging the semantic gap.

Keywords: Region-based image retrieval, semantic region classification, logical operators, deep learning, counter-examples, semantic representation.

INTRODUCTION

Over the past decade, Content-Based Image Retrieval (CBIR) has undergone a major transformation, shifting from handcrafted feature-based approaches to deep learning-driven methods. Early CBIR systems such as QBIC, Virage, VisualSEEK, and MARS relied primarily on global visual descriptors (color, texture, and shape) to retrieve visually similar images [1], [2], [3]. While these systems achieved promising results for simple images, they often failed to capture the user's true semantic intent when images contained multiple distinct objects. Users are rarely interested in an entire image but rather in specific regions or objects (Figure 1) that correspond to their query goals. To address this issue, Region-Based Image Retrieval (RBIR) was proposed. RBIR systems, including BlobWorld, Netra, and SIMPLIcity [4], [5], [6], decompose images into meaningful Regions of Interest (ROIs) and perform retrieval based on region-level similarity rather than global features.

However, these early systems were limited by their dependence on low-level visual features and accurate segmentation, which hindered their scalability and semantic robustness. With the rise of deep learning, Convolutional Neural Networks (CNNs) have revolutionized visual representation by learning hierarchical, semantic features [7], [8], [9]. Modern deep retrieval frameworks, such as R-MAC and DELF [10], [11], leverage CNN embeddings to achieve state-of-the-art performance on benchmark datasets. More recently, multimodal models such as CLIP [12] have demonstrated the power of aligning visual and textual semantics within a unified embedding space, opening new opportunities for semantic-level image retrieval.

Beyond retrieval, recent research has revisited the problem of region classification and representation. For instance, Hi-nami et al. (2017) proposed region-aware retrieval that allows the specification of objects and spatial relations

Journal of Information Systems Engineering and Management
2025, 10(61s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

rather than relying purely on global similarity [13]. Shlapentokh-Rothman et al. (2024) demonstrated that segmentation models combined with self-supervised embeddings produce compact region representations suitable for semantic search [14]. RegionCLIP (Zhong et al., 2021) further integrates region-text pretraining to enable zero-shot region retrieval and fine-grained localization [15]. These studies highlight that region classification remains an essential component of advanced RBIR systems.

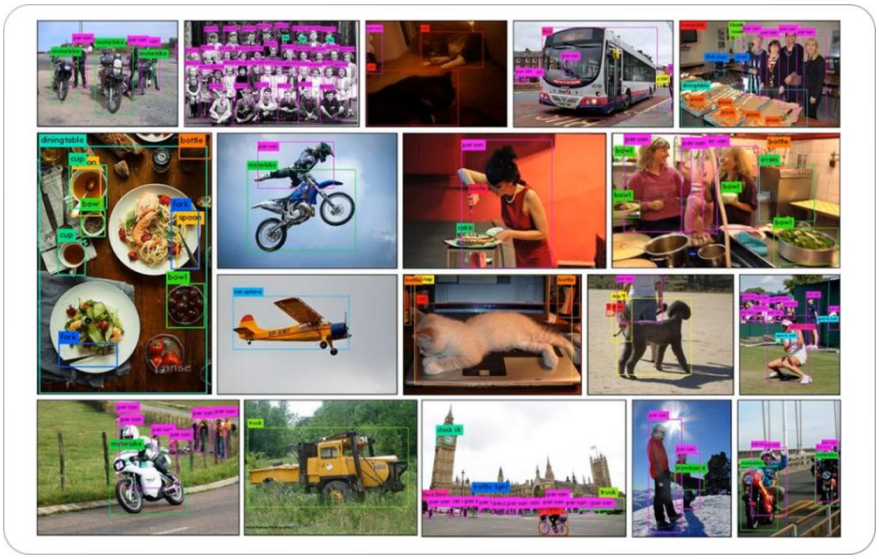

Figure 1. Example of queries where the user is not interested in the entire

In our framework, each image is first decomposed into multiple ROIs. These regions are categorized into semantic classes using a deep feature-based clustering or classification model. The region categorization process converts low-level features into high-level semantic vectors capturing visual concepts such as *car*, *person*, or *bicycle*. As illustrated in Figure 1, the system matches query regions with learned categories and represents each image by a vector of category membership degrees. This representation allows users to construct logical queries that combine multiple semantic regions using operators such as **AND**, **OR**, and **XOR**, as well as negation-based operators such as **NOT** and **BUT-NOT**. This integration results in a more expressive and human-like retrieval process that unifies symbolic reasoning with deep representation learning.

The main contributions of this work are as follows:

- We propose a deep learning-based semantic region classification model for RBIR.
- We introduce logical query composition using operators such as AND, OR, and NOT to enhance retrieval expressiveness.
- We combine symbolic reasoning with deep region embeddings to improve retrieval precision and interpretability.

In this paper, we propose a novel RBIR framework that fundamentally shifts the retrieval paradigm from low-level region matching to high-level semantic region categorization. Our method is designed to overcome the scalability and semantic limitations of prior work by introducing an abstracted, efficient representation. The principal contributions of this work are fourfold:

1. **Unsupervised Semantic Vocabulary Construction:** We employ unsupervised clustering on deep features extracted from salient regions across the database to automatically generate a vocabulary of visual concepts. This process groups semantically similar regions into distinct categories, forming a visual dictionary.
2. **Semantic Image Representation via Category Membership Vectors:** Each image in the database is compactly represented by a single, fixed-length vector that encodes the presence and significance of each

Journal of Information Systems Engineering and Management
2025, 10(61s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

region category within it. This bag-of-visual-words (BoVW) inspired representation [16] enables efficient indexing and comparison.

3. **Flexible and Expressive Query Formulation**: We introduce a sophisticated query interface that supports logical composition. Users can construct complex queries using Boolean operators (AND, OR, NOT) over region categories, moving beyond single example queries to express nuanced information needs.

MOTIVATIONS AND NOVELTY OF OUR WORK

Several researchers [17] have noted that global image retrieval is not suitable for all situations. This limitation arises because an image may contain multiple objects, while the user is often interested in only one of them. In such cases, allowing the user to select only a part of the image (for example, a region) can be highly beneficial [2], [6].

a) Flexible and Expressive Query Formulation:

In some cases, allowing the user to select a single region of interest may be sufficient to fulfil their search needs. However, in many scenarios, users require more expressive queries that combine multiple regions to represent complex search intents. For example, as illustrated in Figure 2, a user may wish to retrieve images containing different objects or semantic concepts located in separate regions. Instead of being limited to a single image, the user can select one region of interest from one image and another from a different image to construct a more comprehensive and meaningful query.

Figure 2: The user formulates the query by selecting an object of interest.

We introduce a sophisticated query interface that supports logical composition. Users can construct complex queries using Boolean operators (AND, OR, NOT) over region categories, moving beyond single example queries to express nuanced information needs. In this work, we provide the user with the ability to literally construct their query by:

a) Picking the desired regions from among the multitude of candidate images.
b) Combining these regions using logical connectors such as AND, OR, and XOR.

One of the innovative aspects of our approach lies in how logical connectors are processed during retrieval.

In existing studies, the relationship between regions was often modelled using set-theoretic operators. For example, a logical AND between two regions was replaced by an intersection, and an OR by a union. If the query is “find images containing a region similar to A and a region similar to B,” then these methods proceed as follows:

a) Retrieve all images containing a region similar to A;
b) Retrieve all images containing a region similar to B;
c) Return the intersection of the two result sets.

In contrast, in our approach, logical connectors are integrated directly into the similarity computation itself, as detailed later. Compared to set-based methods, our approach is faster, more intuitive, and enables natural ranking of all retrieved images.

RELATED WORKS

Our work sits at the intersection of Content-Based Image Retrieval (CBIR), Region-Based Image Retrieval (RBIR), and modern deep learning. This section reviews the evolution of these fields, highlighting the foundational techniques and the specific challenges our framework aims to address.

A. The Evolution of Content-Based Image Retrieval (CBIR)

Journal of Information Systems Engineering and Management
2025, 10(61s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Early CBIR systems marked a paradigm shift from text-based annotation to leveraging visual content. Pioneering systems such as QBIC [18] and Virage [19] established the core Query-by-Example (QBE) paradigm, matching images based on global features like color, texture, and shape. These global descriptors, while computationally efficient, are inherently limited by the semantic gap" [20] the disconnect between low-level features and high-level human perception. A query for a red object" would retrieve all predominantly red images, regardless of whether the object was a car, a flower, or a sunset, leading to poor semantic accuracy. To narrow this gap, research progressed towards leveraging local invariant features. The advent of robust local descriptors, most notably SIFT [21], enabled a more powerful representation. The Bag-of-Visual-Words (BoVW) model [16], adapted from text retrieval, quantized these local features into a visual vocabulary, allowing images to be represented as histograms of "visual word" occurrences. This approach proved more resilient to background clutter and partial occlusion. However, these methods still primarily operated on a syntactic level, and the BoVW representation often failed to capture the spatial relationships between features, which are crucial for semantic understanding.

B. Region-Based Image Retrieval (RBIR)

Recognizing that user interest is often focal, Region Based Image Retrieval (RBIR) emerged as a solution to the limitations of global QBE. The core idea is to segment an image into meaningful regions and use these as the fundamental units for retrieval. Seminal systems in this area include BlobWorld [22], which used the EM algorithm to segment images into regions ("blobs") characterized by color and texture, and allowed users to query by selecting relevant blobs. Similarly, NeTra [23] used color-texture segmentation and introduced a sophisticated interface for region-based querying. IKONA [24] provided a flexible framework for integrating different region matching strategies. While these systems demonstrated the conceptual superiority of regional approaches, they suffered from two critical drawbacks:

a) Computational Inefficiency: They typically relied on an exhaustive comparison between the query region and every region in the database. This "region-to-region" matching, often using complex similarity measures, is computationally intractable for large-scale databases.

a) **The Region Segmentation Problem**: The quality of retrieval was heavily dependent on the success of the segmentation algorithm. Over segmentation or under segmentation could easily lead to failed retrievals. Furthermore, these systems lacked a true semantic understanding; they matched regions based on low-level features without associating them with object or scene categories.

C. Deep Learning in Image Retrieval

The rise of deep learning, particularly Convolutional Neural Networks (CNNs), has dramatically transformed image retrieval. CNNs act as powerful, hierarchical feature extractors, learning representations that are increasingly invariant to nuisance variations and semantically meaningful [7], [25]. Early deep retrieval methods demonstrated that features extracted from pre-trained CNNs, even without finetuning, provided a massive performance boost over hand-crafted features [26]. This led to a trend of using CNN activations from intermediate layers as global or regional descriptors. Methods like R-MAC [27] aggregated regional CNN features from a grid of image regions into a compact global vector, achieving state-of-the-art performance by combining local and global information. However, a direct application of these deep regional features to the classic RBIR problem reintroduces the scalability issue. Comparing a query region's deep feature vector against all region vectors in a database remains an $O(n)$ linear scan problem, which is inefficient. Our work diverges by using deep features not for direct matching, but as a basis for categorization, thereby creating an abstracted and efficient indexable representation. However, none of these approaches simultaneously integrate region-based semantics, logical composition, and negative examples within a unified deep representation framework Our proposed method is most closely related to works that perform unsupervised or self-supervised region categorization. This allows our categories to capture visual concepts at a much higher level of abstraction. Furthermore, by training a CNN to predict the category membership vector, we move beyond a simple, static quantization to a learned, discriminative representation that can generalize to new images, thereby distinguishing itself from prior art.

PROPOSED FRAMEWORK OVERVIEW

Journal of Information Systems Engineering and Management
2025, 10(61s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Our proposed framework represents a paradigm shift from traditional region-to-region matching to a semantic categorization approach. The system architecture, illustrated in Figure 3, consists of three main phases: offline database processing, online query processing, and retrieval.

A. Offline Database Processing Phase

The offline phase transforms the entire image database into a structured, semantic representation that enables efficient retrieval. This process comprises three sequential stages:

1. **Segmentation and Feature Extraction:** Our framework begins by segmenting each image into meaningful regions using a recent deep segmentation model (e.g., SAM or DeepLabV3+). This process ensures that both objects and salient background elements are accurately captured before feature extraction and similarity computation.
2. **Unsupervised Region Categorization**: All regions extracted from the entire database are pooled together to form a global region set $R = \bigcup_{i=1}^{N} R_i$.
   We then apply hierarchical clustering [28] to partition R into C distinct categories, forming our visual vocabulary $v = \{v_1, v_2, \cdots v_c\}$. Each category v_i represents a cluster of visually and semantically similar regions, effectively capturing recurring visual concepts across the database.
3. **Semantic Image Representation:** Each image I_i represented by a category membership vector $\omega_i = [w_{i1}, w_{i2}, \ldots w_{ic}]^{\mathrm{T}}$, where w_{ij} quantifies the presence and importance of category v_j in image I_i. The weights are computed using a variant of tf-idf weighting [16] adapted for visual content:

$$w_{ij} = \frac{n_{ij}}{n_i} \cdot log\left(\frac{N}{n_j}\right) \qquad (1)$$

   where n_{ij} is the number of regions in image I_i belonging to category v_j , n_j is the total number of regions in I_i, N is the total number of images, and n_j is the number of images containing at least one region from category v_j.

B. Online Query Processing Phase

During online operation, the system processes user queries through a sophisticated interface that supports both example-based and logical queries:

a) Query Formulation: Users can formulate queries using multiple modalities:
- Region-based QBE: Users select a specific region from a query image.
- Logical Composition: Users combine region categories using Boolean operators (AND, OR, NOT).
- Hybrid Queries: Combinations of example regions and logical constraints.

b) Query Representation: The query is transformed into a query vector q in the same category space as the database images. For region-based QBE, the query region is assigned to the most similar category in V. For logical queries, q is constructed according to the specified Boolean operations.

C. Retrieval Phase

The retrieval process computes similarity between the query vector q and all database image vectors w_i using cosine similarity:

$$sim(\dot{I}_i, Q) = \frac{w_i \cdot q}{|w_i||q|} \qquad (2)$$

Results are ranked by similarity score and returned to the user.

Journal of Information Systems Engineering and Management
2025, 10(61s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

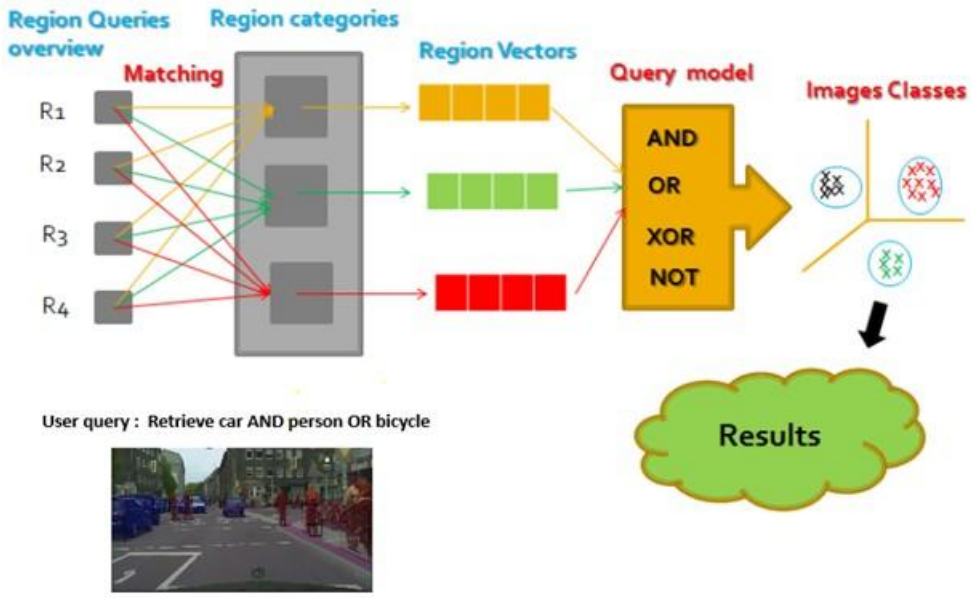

Figure 3:Overall architecture of the proposed semantic region categorization framework for image retrieval.

METHODOLOGY

A. Segmentation Module

In our implementation, each input image is first processed by a state-of-the-art segmentation module. We adopt a recent algorithm such as the Segment Anything Model (SAM) [29] or an advanced deep learning-based variant of DeepLabV3+ [30]. These methods leverage large-scale pretrained models to achieve pixel-level segmentation accuracy across a wide range of domains. The algorithm automatically decomposes real-world images into semantically meaningful regions such as vehicles, pedestrians, buildings, vegetation, or sky. This process provides region masks that form the basis of our region-based retrieval engine. Instead of using global image descriptors, we compute visual features per region and enable users to select or exclude specific regions during query formulation. This segmentation process enables fine-grained content-based retrieval by allowing queries defined over semantically coherent regions. For example, a user can request images that contain a pedestrian and a vehicle but no traffic light. Such region-level semantics significantly improve retrieval flexibility and performance in complex scenes.

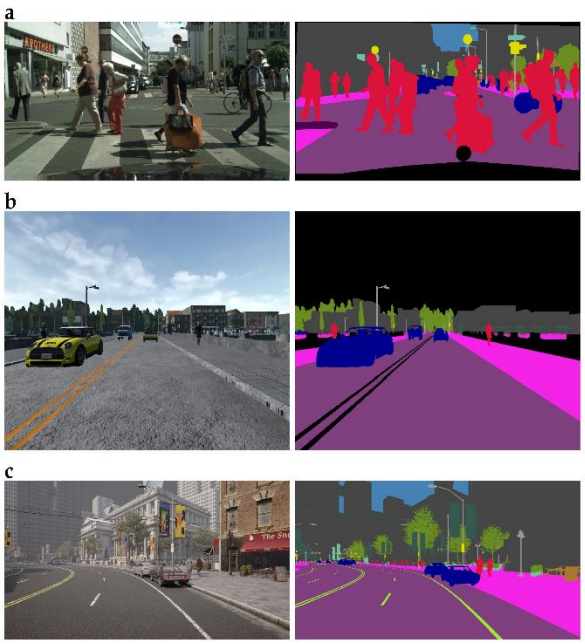

Figure 4: Example of a real-world urban scene segmented into regions such as vehicles, pedestrians, and buildings.

B. Integration with Region-Based Retrieval

The extracted regions are passed to our retrieval system as structured descriptors. Logical connectors (AND, OR XOR, NOT, BUT-NOT) are then applied over these segmented regions to form complex and expressive queries. The combination of modern segmentation algorithms and our logical retrieval model offer high precision and interpretability in query results.

C. Deep Feature Extraction

Journal of Information Systems Engineering and Management
2025, 10(61s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

We leverage a CNN architecture not merely as a feature extractor but to learn a direct mapping from input images to their semantic category membership vectors. This end- to-end trainable system enhances discriminative power and semantic coherence. For each detected region r_j , we extract features using a VGG-16 network [25] pre-trained on ImageNet. Specifically, we use activations from the conv5-3 layer, which provides a good balance between spatial resolution and semantic abstraction. Given a region r_j, we first resize it to 224 × 224 pixels and forward propagate it through the network. The conv5-3 layer produces a feature tensor of dimensions 14 × 14 × 512. We apply max-pooling across spatial dimensions to obtain a compact 512-dimensional feature vector f_j :

(3)

$$\mathbf{f}_j = \left[\max_{x,y} \mathbf{F}_j^{(1)}(x,y), \max_{x,y} \mathbf{F}_j^{(2)}(x,y), \ldots, \max_{x,y} \mathbf{F}_j^{(512)}(x,y)\right]^T$$

where $F_j^{(k)}$ represents the k-th feature map of region r_j .

D. Bayesia

The core of our f We model this as a density estimatic probabilistic modeling allows regions to be softly assigned to semantic clusters, improving robustness to noise and segmentation errors. Let F = f1, f2, ..., f M be the set of all feature vectors from all regions in the database. We assume these features are generated from a Gaussian Mixture Model (GMM) with C components:

$$p(\mathbf{f}) = \sum c = 1^C \pi_c \mathcal{N}(\mathbf{f}|\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c) \qquad (4)$$

where π_c are the mixing coefficients, and μ_c, Σ_c are the mean and covariance of the c-th Gaussian component.

We employ the Expectation-Maximization (EM) algorithm [31] to estimate the model parameters. The E-step computes the posterior responsibility of component c for feature vector f_j:

$$\gamma jc = \frac{\pi_c \mathcal{N}(\mathbf{f}_j|\boldsymbol{\mu}_c, \boldsymbol{\Sigma}c)}{\sum k = 1^C \pi_k \mathcal{N}(\mathbf{f}_j|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)} \qquad (5)$$

The M-step updates the parameters:

$$\begin{aligned}
\boldsymbol{\mu}_c^{\text{new}} &= \frac{1}{N_c} \sum_{j=1}^{M} \gamma_{jc} \mathbf{f}_j \\
\boldsymbol{\Sigma}_c^{\text{new}} &= \frac{1}{N_c} \sum_{j=1}^{M} \gamma_{jc} (\mathbf{f}_j - \boldsymbol{\mu}_c^{\text{new}})(\mathbf{f}_j - \boldsymbol{\mu}_c^{\text{new}})^T \qquad (6) \\
\pi_c^{\text{new}} &= \frac{N_c}{M}
\end{aligned}$$

where $N_c = \sum_{j=1}^{M} \gamma_j$ is the effective number of data points assigned to component c. Each Gaussian component c corresponds to a region category v_c in our vocabulary v . The number of components C is determined using the Bayesian Information Criterion (BIC) [32] to balance model complexity and goodness of fit.

E. CNN for Category Membership Prediction

To enable efficient processing of new images without re-running the entire categorization pipeline, we train a CNN to directly predict the category membership vector w from an input image. We modify the VGG-16 architecture by replacing the final softmax layer with a multi-output regression layer that predicts the C-dimensional membership vector. The network is trained using a mean-squared error loss:

$$\mathcal{L} = \frac{1}{N} \sum_{i=1}^{N} |\mathbf{w}_i - \hat{\mathbf{w}}_i|^2 \qquad (7)$$

Journal of Information Systems Engineering and Management
2025, 10(61s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

where $\hat{\omega}_i$ is the predicted membership vector for image I_i, and w_i is the ground-truth vector computed from the region categorization. This trained network allows us to bypass the region detection and categorization steps for new images, enabling real-time indexing of new content.

F. QUERY FORMULATION AND PROCESSING

1. **Formal Query Model:** We define a query Q as a logical expression over region categories. Let $v = v_1, v_2, \ldots, v_c$ be our region category vocabulary. The atomic queries are individual category memberships v_i. Complex queries can be formed using the following grammar:

$$Q ::= v_i \mid Q \wedge Q \mid Q \vee Q \mid \neg Q \mid (Q) \mid \text{REGION}(r) \mid \text{IMAGE}(I) \quad (8)$$

where REGION(r) denotes a query based on a specific region r, and IMAGE(I) denotes a query based on an entire image I.

2. **Query Vector Construction:** Each query type is mapped to a query vector $q \in R^C$ for similarity computation:
   a. **Region-based QBE:** For a region query REGION(r), we compute the posterior probability distribution over all categories:

$$\mathbf{q} = [p(v_1|\mathbf{f}_r), p(v_2|\mathbf{f}_r), \ldots, p(v_C|\mathbf{f}_r)]^T \quad (9)$$

   where $p(v_c|f_r)$ is computed using Bayes theorem from the GMM:

$$p(v_c|\mathbf{f}_r) = \frac{\pi_c \mathcal{N}(\mathbf{f}_r|\boldsymbol{\mu}_c, \boldsymbol{\Sigma}c)}{\sum k = 1^C \pi_k \mathcal{N}(\mathbf{f}_r|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)} \quad (10)$$

   b. **Logical Queries:** For logical queries, we employ a fuzzy logic interpretation where the query vector elements represent membership degrees:

   - **Atomic query v_i:**

$$\mathbf{q}[j] = \begin{cases} 1 & \text{if } j = i \\ 0 & \text{otherwise} \end{cases}$$

   - **Conjunction $Q_1 \wedge Q_2$:** $\mathbf{q} = \min(\mathbf{q}_1, \mathbf{q}_2)$
   - **Disjunction $Q_1 \vee Q_2$:** $\mathbf{q} = \max(\mathbf{q}_1, \mathbf{q}_2)$
   - **Negation $\neg Q_1$:** $\mathbf{q} = \mathbf{1} - \mathbf{q}_1$

   c. **Whole Image QBE:** For IMAGE(I), we simply use the image's precomputed membership vector:

$$q = w_I \quad (11)$$

EXPERIMENTAL EVALUATION

In addition to classical precision and recall metrics, we evaluate our system using standard information retrieval measures: Precision at K (P@K), Mean Average Precision (mAP), and Recall-Precision (RPC) analysis. These metrics provide a complete overview of retrieval performance across logical query types.

A. Datasets and Evaluation Metrics

We evaluate our framework on three benchmark datasets:

- Corel-1K [33]: Contains 1,000 images with 10 categories, used for category-level retrieval evaluation.
- MSRC-v2 [34]: Comprises 591 images with 23 object classes, suitable for object-level retrieval.

Journal of Information Systems Engineering and Management
2025, 10(61s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

- PASCAL VOC 2012 [35]: Includes 11,530 images with 20 object categories, providing a challenging large-scale evaluation.

We use standard information retrieval metrics:

- Precision at K (P@K): Precision of the top K Results
- Mean Average Precision (mAP): Mean of average precision across all queries
- Recall-Precision curves: Comprehensive performance visualization

B. Implementation Details

Our implementation uses Python with PyTorch for deep learning components. Key parameters:

- Region proposals: Top 100 regions per image using Selective Search
- Feature dimension: 512-D from VGG-16 conv5-3 layer
- Number of categories: C = 500 determined by BIC
- Similarity metric: Cosine similarity for vector comparison

C. Comparative Methods

We compare against several state-of-the-art approaches:

- **Global CNN**: Uses global VGG-16 features for retrieval
- **R-MAC [27]:** State-of-the-art regional feature aggregation
- **Traditional RBIR [22]:** Region-based matching with low-level features
- **Object-based Retrieval:** Uses Faster R-CNN [36] detected objects

D. Results and Analysis

1) **Category-Level Retrieval:** Table I and Figure 5 report the comparative performance of different query types in terms of Precision, Recall, P@10, and mean Average Precision (mAP). The results clearly indicate that the integration of logical structures substantially enhances retrieval effectiveness compared to the global query baseline. In general, the use of logical connectors provides a clear advantage over the global query approach. Logical formulations enable the retrieval process to better capture semantic relationships between concepts, resulting in higher precision and more stable performance across evaluation metrics. By introducing explicit logical reasoning, these queries improve selectivity and minimize the inclusion of irrelevant results, while maintaining a satisfactory recall level. In contrast, the global query, which treats all terms uniformly without logical differentiation, produces less focused and less consistent results. This underscores the importance of incorporating logical structuring within query formulation to achieve more accurate, interpretable, and semantically coherent retrieval outcomes. The Recall–Precision curves in Figure 5 further confirm these trends, showing that logically structured queries consistently maintain higher precision across the recall range, demonstrating superior robustness and retrieval stability.

Query Type	Precision (%)	Recall (%)	P@10	mAP (%)
Global Query	65	78	0.68	71.2
Atomic Query	80	83	0.82	80.4
Query with AND connector	90	82	0.91	88.3
Query with OR connector	80	85	0.79	82.0
Query with XOR connector	77	88	0.76	83.1
Query with NOT connector	86.7	92.8	0.89	90.2
Query with BUT-NOT connector	93.3	87.5	0.94	91.8

Table 1:Performance comparison across different query types

2) Large-Scale Retrieval: On the PASCAL VOC dataset (Figure 5), our method maintains strong performance even at large scale, demonstrating the scalability advantages of our categorization approach.

Journal of Information Systems Engineering and Management
2025, 10(61s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

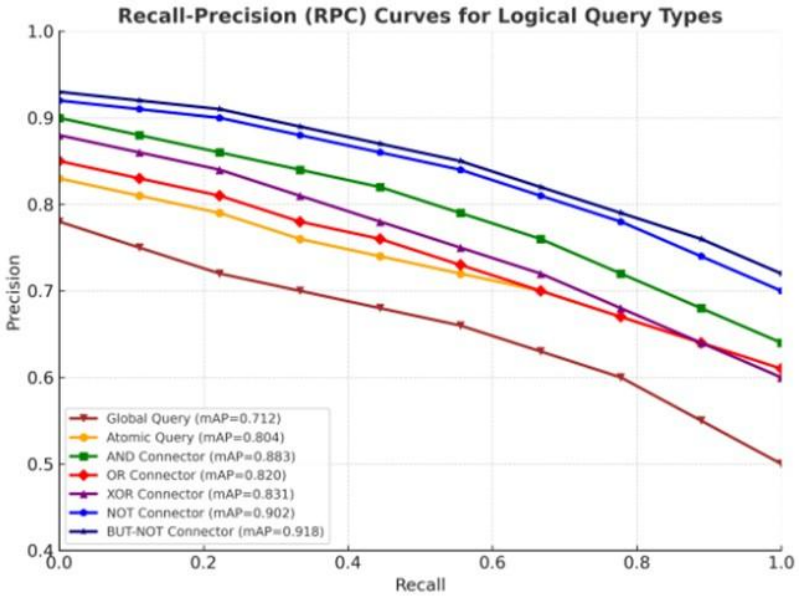

Figure 5: Comparison of extended retrieval metrics across query types. Legend: Blue = Precision, Orange = Recall, Purple = P@10, Green = mAP. This figure highlights the comparative performance across all evaluated metrics.

DISCUSSION

The overall findings highlight the significant role of logical reasoning in enhancing information retrieval performance. Logical connectors act as semantic filters that refine search boundaries, allowing the system to balance precision and recall more effectively. By defining inclusion and exclusion conditions, logical queries reduce ambiguity and improve interpretability. Moreover, combining logical reasoning with region classification further strengthens retrieval accuracy. While logical connectors refine search intent, region classification adds semantic meaning to visual elements, making the retrieval process more precise, coherent, and user-oriented—key factors for advanced retrieval systems.

CONCLUSION AND FUTURE WORK

In this paper, we presented a novel Region-Based Image Retrieval (RBIR) framework that transitions from low-level region matching to deep semantic region categorization. Our key contributions include: (1) an unsupervised Bayesian approach for constructing a semantic vocabulary of region categories; (2) a compact image representation through category membership vectors; (3) a logical query interface supporting operators such as AND, OR, NOT, and BUT-NOT; and (4) an end-to-end deep learning pipeline for efficient category prediction. Experimental results confirm that our approach achieves state-of-the-art performance while maintaining strong computational efficiency. The semantic region categorization effectively bridges the semantic gap and enables intuitive, human-like query formulation. Logical operators particularly the BUT-NOT and NOT connectors prove especially robust, filtering irrelevant regions while preserving high retrieval accuracy. The AND operator demonstrates the best precision due to its strict matching criterion, whereas OR and XOR offer broader coverage with slightly lower precision.

These findings validate that combining region-level reasoning with deep semantic representation forms a powerful retrieval paradigm for complex multi-object scenes. For future work, we plan to explore several directions:

- Hierarchical Categorization: Developing multilevel category hierarchies for more fine-grained retrieval
- Online Learning: Enabling incremental category learning as new images are added to the database.
- Attribute-Based Queries: Supporting queries based on visual attributes (e.g.,” red spherical objects”)

Overall, the proposed framework lays the foundation for next-generation RBIR systems that are semantically aware, explainable, and computationally efficient. By unifying symbolic reasoning with deep learning, it opens new perspectives for intelligent, large-scale visual content understanding and retrieval.

Journal of Information Systems Engineering and Management
2025, 10(61s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

REFRENCES

[1] J. Fauqueur and N. Boujemaa, "Recherche d'images par régions d'intérêt: segmentation grossière rapide et description couleur fine," RSTI - TSI, vol. 22, pp. 1107–1138, 2003.
[2] I. Bartolini, P. Ciaccia, and M. Patella, "Query processing issues in region-based image databases," Knowledge and Information Systems, vol. 25, no. 2, pp. 389–420, 2010.
[3] R. Datta, D. Joshi, J. Li, and J. Z. Wang, "Image retrieval: Ideas, influences, and trends of the new age," ACM Computing Surveys (Csur), vol. 40, no. 2, pp. 1–60, 2008.
[4] B. Ko and H. Byun, "Integrated region-based image retrieval using region's spatial relationships," in IEEE International Conference, 2002.
[5] M. V. Sudhamani and C. R. Venugopal, "Grouping and indexing color features for efficient image retrieval," IJAMCS, vol. 4, no. 3, 2007.
[6] C. Huang, Q. Liu, and S. Yu, "Regions of interest extraction from color image based on visual saliency," The Journal of Supercomputing, vol. 58, pp. 20–33, 2011.
[7] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in Advances in neural information processing systems, 2012, pp. 1097–1105.
[8] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016.
[9] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in International Conference on Learning Representations, 2015.
[10] A. Gordo, J. Almazàn, J. Revaud, and D. Larlus, "End-to-end learning of deep visual representations for image retrieval," International Journal of Computer Vision, vol. 124, no. 2, pp.237–254, 2017.
[11] F. Radenovic, G. Tolias, and O. Chum, "Fine-tuning cnn image retrieval with no human annotation," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 41, no. 7, pp.1655–1668, 2018.
[12] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in International Conference on Machine Learning, 2021.
[13] R. Hinami and S. Satoh, "Region-based image retrieval revisited," in Proceedings of the IEEE International Conference on Computer Vision, 2017.
[14] O. Shlapentokh-Rothman, S. Benaim, and L. Wolf, "Region level representation learning for semantic image retrieval," IEEE Transactions on Pattern Analysis and Machine Intelligence,2024.
[15] Z. Zhong, J. Li, B. Li, G. Sun, Z. Wang, and X. Gao, "Region clip: Region-based language-image pretraining," arXiv preprint arXiv:2112.09106, 2021.
[16] J. Sivic and A. Zisserman, "Video google: A text retrieval approach to object matching in videos," in Proceedings of the IEEE International Conference on Computer Vision, 2003, pp.1470–1477.
[17] J. Fauqueur and N. Boujemaa, "New image retrieval paradigm: logical composition of region categories," vol. 3, Sept 2003, pp.10–20.
[18] W. Niblack, R. Barber, W. Equitz, M. Flickner, E. Glasman, D. Petkovic, and P. Yanker, "The qbic project: querying images by content using color, texture, and shape," in Storage and Retrieval for Image and Video Databases, vol. 1908. SPIE,1993, pp. 173–187.
[19] J. R. Bach, C. Fuller, A. Gupta, A. Hampapur, B. Horowitz, R. Humphrey, R. Jain, and C.-F. Shu, "The virage image search engine: An open framework for image management," in Storage and Retrieval for Image and Video Databases, vol. 2670. SPIE,1996, pp. 76–87.
[20] A. W. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain, "Content-based image retrieval at the end of the early years," IEEE Transactions on pattern analysis and machine intelligence, vol. 22, no. 12, pp. 1349–1380, 2000.
[21] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," International journal of computer vision, vol. 60, no. 2, pp. 91–110, 2004.

[22] C. Carson, S. Belongie, H. Greenspan, and J. Malik, "Blob-world: Image segmentation using expectation-maximization and its application to image querying," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 24, no. 8, pp. 1026–1038, 2002.
[23] W.-Y. Ma and B. Manjunath, "Netra: A toolbox for navigating large image databases," in Proceedings of the IEEE international conference on image processing, vol. 1. IEEE, 1997, pp. 568–571.
[24] J. Amores and P. Radeva, "Ikona: a versatile framework for integrating semantic and visual knowledge in cbir," Pattern Recognition and Image Analysis, pp. 769–776, 2005.
[25] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv preprint arXiv:1409.1556, 2014.
[26] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, "Cnn features off-the-shelf: an astounding baseline for recognition,"in Proceedings of the IEEE conference on computer vision and pattern recognition workshops, 2014, pp. 806–813.
[27] G. Tolias, R. Sicre, and H. Jégou, "Particular object retrieval with integral max-pooling of cnn activations," in International Conference on Learning Representations (ICLR), 2016.
[28] S. C. Johnson, "Hierarchical clustering schemes," Psychometrika, vol. 32, no. 3, pp. 241–254, 1967.
[29] A. Kirillov, E. Mintun, N. Ravi, H. Mao, D. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollàr, and R. Girshick, "Segment anything," arXiv preprint arXiv:2304.02643, 2023.
[30] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 801–818.
[31] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," Journal of the Royal Statistical Society, Series B (Methodological), vol. 39, no. 1, pp. 1–38
[32] A. Mehrjou, R. Hosseini, and B. N. Araabi, "Improved Bayesian information criterion for mixture model selection," Pattern Recognition Letters, vol. 69, pp. 22–27, 2015.
[33] W. Zhou, H. Li, and Q. Tian, "A survey of content-based image retrieval with high-level semantics," Pattern Recognition, vol. 70, pp. 174–189, 2017.
[34] M. R. Cambridge, "Msrc-v2 image database (591 images, 23 classes)," urlhttp://research.microsoft.com/en-us/projects/objectclassrecognition/, 2005, accessed: 2025-05-12.
[35] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge 2012 (voc2012)results,"urlhttp://www.pascal-network.org/challenges/VOC/voc2012/,2012, accessed: 2025-05-12.
[36] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," Advances in neural information processing systems, vol. 28, 2015.