A MULTI-SOURCE CBIR APPROACH LEVERAGING AUTOENCODERS FOR IMAGE ORGANIZATION AND DEEP LEARNING FOR SURROGATE LABELING

Manish Rai and Rupali Sharma

ABSTRACT:

Deep learning-based Content-Based Image Retrieval (CBIR) has emerged as a leading research area due to its ability to deliver more accurate search results than traditional methods, despite being computationally demanding. This study presents an innovative approach that enhances CBIR performance through a multilevel aggregation technique integrated with autoencoders, aimed at precise feature selection and improved search accuracy. The proposed method employs a dual strategy. First, a multilevel aggregation technique is used to process image features at various granularities, capturing both local and global image attributes for a comprehensive representation. Autoencoders then reduce the dimensionality of these features, retaining only the most relevant information while reducing computational costs. The refined features are leveraged to effectively tag the search key, guiding the search process toward the most pertinent target images. Second, the method distinguishes between locally significant datasets and generic ones, employing specific approaches for each. For domain-specific datasets, unique characteristics are utilized to enhance retrieval rates, while a more generalized strategy is applied to broader datasets. Query expansion is implemented to broaden the search scope, incorporating additional relevant terms or images related to the original query. Additionally, pseudo-labeling is introduced, where deep learning models classify images into positive (similar to the query) and negative classes. Query image features are compared to those in the search pool using assigned weights, and target images are ranked based on an adaptive threshold. Tested on public datasets, this method demonstrates significant improvements in precision, recall, and computational efficiency compared to recent approaches. This robust technique shows promise for applications requiring precise image retrieval, such as medical imaging, security surveillance, and digital asset management.

Index Terms: *Deep learning; CBIR; review; image retrieval.*

Reference to this paper should be made as follows:

Manish Rai and Rupali Sharma, (2024), "A Multi-Source CBIR Approach Leveraging Autoencoders for Image Organization and Deep Learning for Surrogate Labeling", Vol. 12, No. 3, pp. 01-18, DOI 10.30696/IJEEA.XII.III.2024.01-18

Biographical notes:

Manish Rai *is Asst. Professor in S.S.M Engineering and Management College, Karnataka, India.*

Rupali Sharma *PhD Scholar at Karnataka University, Dharwad, Karnataka, India.*

1. INTRODUCTION

The rapid increase in digital devices and social media usage has led to an unprecedented growth in the volume of digital content generated and stored. This surge has made it crucial to develop advanced storage systems and efficient methods for retrieving stored content on demand. Consequently, the expansion of cyberspace has sparked extensive research into managing digital data profiles and exploring effective solutions for image retrieval without relying on textual annotations. This has given rise to Content-Based Image Retrieval (CBIR), a widely adopted method that uses computer vision techniques to retrieve digital images relevant to a query based on their visual content.

Unlike traditional Text-Based Image Retrieval (TBIR) systems, which rely on matching textual information associated with images, CBIR analyzes intrinsic image features such as color, texture, shape, and spatial relationships to find visually similar images. This makes CBIR a powerful tool, especially in situations where images lack descriptive text or metadata. While TBIR is limited to images with textual descriptors, CBIR is more versatile and can be applied in various fields, including search engines, medical imaging, digital libraries, and the tourism industry. For instance, in medical applications, CBIR helps find visually similar medical images for diagnosis and research, whereas in tourism, it can assist users in discovering visually similar landmarks or locations.

CBIR systems support different query types, including query by text, where textual descriptions are used; query by image, which utilizes an example image as the search input; query by sketch, where a user's sketch serves as the basis for the search; and query by concept, which involves retrieving images based on abstract ideas. These diverse query methods make CBIR a robust solution for handling various search scenarios.

In summary, CBIR is a significant advancement in image retrieval technology, offering a flexible, efficient, and accurate means of accessing vast image databases. It continues to be a focus of active research and development, with ongoing efforts aimed at enhancing retrieval precision and reducing computational demands.

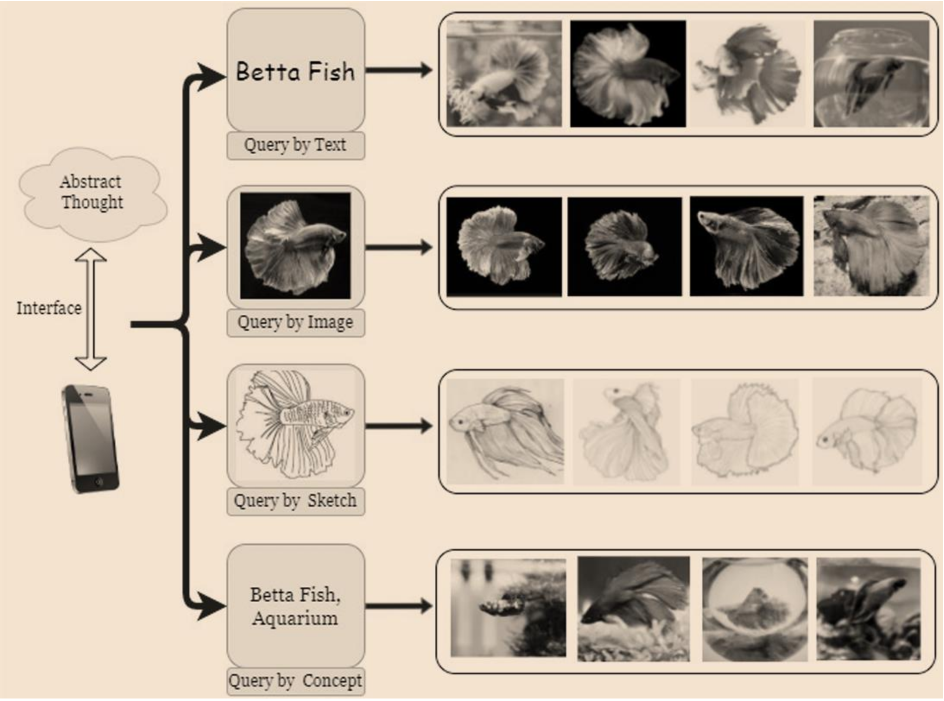


Figure 1: The display of various image retrieval techniques.

1.1 Framework of CBIR

A typical Content-Based Image Retrieval (CBIR) system comprises two essential modules: feature extraction and similarity measurement. In the feature extraction module, the system analyzes an image to identify its key attributes and stores them as a feature vector for future use. This feature vector acts as a compact representation of the image's content, capturing critical visual characteristics such as color, texture, and shape. During the similarity measurement phase, the feature vector of the query image is compared with those of other images in the database to identify similar images. Various similarity measurement techniques are employed to evaluate the degree of resemblance between feature vectors. The images that best match the query are then selected and presented to the user. To ensure efficient and accurate image retrieval, a CBIR framework typically follows this two-stage process of feature extraction and similarity measurement.

Feature extraction is pivotal to the CBIR system's effectiveness, as the quality of the extracted features directly influences the system's ability to accurately represent and differentiate images. This process involves summarizing pixel information from regions of interest within the image to produce a concise, yet informative, representation. The most common features used in CBIR include color, texture, and shape, which together provide a comprehensive depiction of an image's visual properties. Color features are widely used due to their strong visual appeal and ability to represent the spectral information of an image. Texture features capture variations in intensity and are valuable for identifying patterns and surface properties. Shape features describe the geometric structure of objects within the image, focusing on the contours and outlines of specific regions.

Feature extraction can be performed at different granularities, yielding either global or local features. Global features describe the entire image and include descriptors such as overall color distribution, texture, and shape. Local features, on the other hand, are extracted from specific regions or pixels, capturing detailed information about smaller, localized patterns. These low-level features, whether global or local, are critical in enabling the CBIR system to index and retrieve images effectively based on their visual content.

1.2 Challenges of CBIR

The saying "a picture is worth a thousand words" highlights how images can convey meanings and emotions more effectively than words alone. Humans possess an innate ability to analyze and interpret complex visual information, making images a powerful medium for expressing intricate and multifaceted ideas. However, machines, which rely on numerical comparisons for image analysis, often struggle to achieve the same level of understanding as humans. This discrepancy creates a significant semantic gap between human and machine perception of images, as humans can intuitively grasp the context and deeper meanings that machines typically overlook. Consequently, machine-based analysis often falls short in comparison to human interpretation.

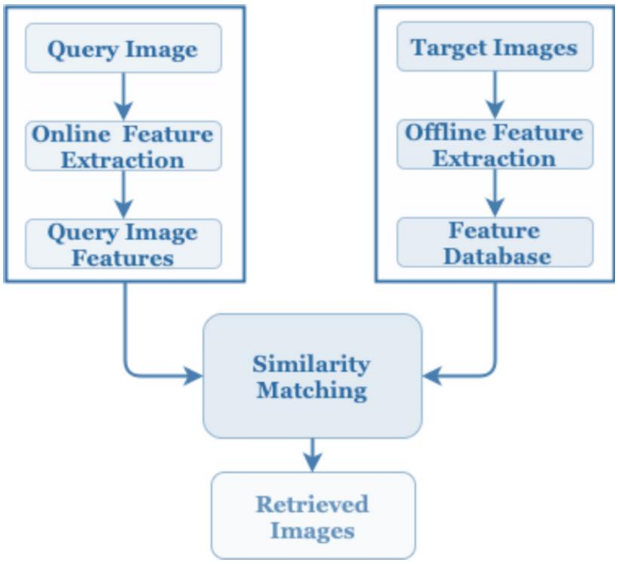


Figure 2: A typical CBIR system's framework.

To bridge this gap, researchers have turned to semantic image retrieval methods enhanced by machine learning. This approach aims to help machines better understand and retrieve images based on their semantic content, rather than just low-level visual features like color, shape, and texture. Retrieving images from large databases poses several challenges. With modern images often having high pixel resolutions, directly comparing them is computationally infeasible. Instead, using low-dimensional features for comparison is more practical. While traditional methods focus on features such as shape and color, they struggle to capture the semantic meaning within an image. This limitation is due to the lack of techniques capable of interpreting the complexities of human perception and contextual understanding.

Another factor contributing to the limitations of traditional image retrieval methods is their non-adaptive nature, which restricts the effective extraction of general image characteristics. This inflexibility hampers the semantic processing of images, even when additional features are mapped. The gap between the data retrieved and the true meaning of an image is known as the semantic gap. The nature of the database is crucial in reducing this gap. When databases are confined to specific domains, the semantic gap is minimized because such databases contain images with limited detail and granularity, making them easier to retrieve accurately.

On the other hand, generic databases, which encompass a wide array of images with varying themes, backgrounds, and contexts, present a greater challenge. The diversity and high dimensionality of these databases make it difficult to accurately detect relevant images. Furthermore, the presence of high-resolution images with diverse contexts complicates the retrieval process. To address these challenges, Smeulders et al. [27] suggested incorporating external information about the images to improve retrieval performance.

Most current Content-Based Image Retrieval (CBIR) systems utilize generic databases, making the reduction of the semantic gap a continuous focus of research. Despite these challenges, advances in machine learning and deep learning are paving the way for more sophisticated semantic understanding and image retrieval capabilities. These developments hold promise for narrowing the semantic gap, thereby enhancing the effectiveness and efficiency of CBIR systems.

1.3 ML and CBIR

Recent advancements in vision-based algorithms have greatly enhanced the precision with which low-dimensional images can be represented, providing substantial benefits to Content-Based Image Retrieval (CBIR) systems. By incorporating machine learning techniques, these systems have improved various stages of image processing, such as noise reduction and feature vector extraction. A well-designed feature vector is critical for accurately representing the essential characteristics of an image, thereby improving the efficiency and effectiveness of image retrieval. Therefore, the development of reliable feature descriptors is fundamental to the success of CBIR systems.

Several widely-used feature descriptors have proven effective in this regard, including Histogram of Oriented Gradients (HOG), Binary Robust Invariant Scalable Keypoints (BRISK), Maximally Stable Extremal Regions (MSER), Scale-Invariant Feature Transform (SIFT), and Speeded-Up Robust Features (SURF) [7, 19, 3, 21]. These descriptors are adept at capturing various elements of an image, such as edges, keypoints, and textures, which are essential for differentiating between visually similar images.

However, CBIR systems still face challenges, including high computational demands and difficulties in managing complex visual data, which can result in suboptimal performance [26]. To overcome these limitations, new decision-making approaches and the application of advanced machine learning methods [35, 34] offer promising opportunities to enhance the overall performance and reliability of CBIR systems. These innovations can lead to more efficient and accurate image retrieval processes, addressing current challenges and expanding the potential applications of CBIR technology.

1.4. Introduction to DL

Machine learning has significantly improved the efficiency of Content-Based Image Retrieval (CBIR) systems by using lower-dimensional feature vectors to represent source data. However, traditional shallow retrieval models often fail to meet expected retrieval rates due to various limitations such as a limited number of training samples, unclear feature spaces, and class imbalance issues. These constraints hinder their ability to capture and retrieve relevant images accurately. In contrast, recent studies have demonstrated that deep learning algorithms can outperform shallow models in generic image retrieval tasks, especially when dealing with large datasets [18, 31, 20, 32, 22].

Deep learning approaches leverage the immense computational power of GPUs, which are not only highly capable but also increasingly affordable. These algorithms excel in image retrieval because they are highly adaptable and can effectively represent complex features. A prominent architecture in this field is the Deep Convolutional Neural Network (CNN) [15]. CNNs consist of multiple layers of convolutions and subsampling, along with activation functions and fully connected layers. They process input images as three-dimensional vectors, taking into account height, width, and depth. As the data passes through the network, it transforms into a single-dimensional vector through fully connected layers, enabling the system to capture intricate image features.

This transformation is particularly valuable in CBIR systems, where a robust feature representation is critical. CNN models can be tailored to improve retrieval performance by adjusting input configurations and classification intervals, especially when trained on extensive datasets. Recent research has also explored the integration of autoencoders with CNNs to enhance image retrieval. Autoencoders, a type of neural network designed to learn efficient data representations, help refine the feature extraction process by reducing dimensionality and eliminating noise, thus improving the overall performance of CBIR systems.

Recognizing the need for more efficient CBIR search techniques, this paper proposes a novel hybrid approach that combines traditional and advanced methods. The proposed CBIR system integrates traditional unsupervised machine learning techniques, such as clustering and autoencoders, with supervised CNN-based approaches. This multi-layered methodology allows the system to explore and filter image contexts from both the query image and the database repository across various levels, resulting in a more nuanced and accurate retrieval process.

The paper is structured as follows: Section 2 provides a detailed review of the state-of-the-art techniques used to enhance retrieval rates in CBIR systems. Section 3 introduces the proposed methodology, explaining the hybrid approach in detail. Section 4 presents the results and discussions, evaluating the performance of the proposed system. Finally, Section 5 concludes the paper, summarizing the findings and suggesting directions for future research in this domain.

2. LITERATURE SURVEY

The Query by Image Content (QBIC) technique [9] is recognized as one of the foundational methods in the first generation of Content-Based Image Retrieval (CBIR) systems. It utilizes fundamental image features such as textures, color palettes, and local objects to retrieve images and video frames. While the simplicity of QBIC, based on direct pixel comparisons, is advantageous, it also has significant limitations. The method is less effective with images that have undergone slight translations or rotations, as its reliance on pixel-based comparisons hampers performance under such variations. This limitation has led researchers to shift focus from pixel-based methods to feature descriptor-based analyses, emphasizing hand-crafted low-dimensional descriptors. This shift has facilitated the development of robust feature extraction techniques and machine learning methods that have substantially improved CBIR capabilities.

Low-dimensional feature descriptors offer several advantages, including reduced error rates, lower computational costs, and increased robustness against geometric distortions like rotation and translation. For instance, Yuan et al. [38] introduced a CBIR method utilizing Local Binary Patterns (LBP) [12] and Scale-Invariant Feature Transform (SIFT), focusing on extracting key points to define the feature space for comparison. This approach involved grouping the local features of target images into clusters, followed by conducting similarity searches on these clustered feature vectors to identify similar images. An enhanced version of this model [37] expanded the feature space by incorporating Histogram of Oriented Gradients (HOG) features, further boosting performance when processing geometrically distorted images. However, this method remains sensitive to noise and can be computationally intensive due to the complexity of managing large feature descriptors like HOG.

In another approach, Yuan et al. [23] proposed a CBIR system that combines color fusion and texture descriptors with the Laplacian score for dimensionality reduction of the feature vector, thereby streamlining the retrieval process while maintaining accuracy. Bibi et al. [4] also made significant contributions by using a set of sparse complementary features for robust image representation and selection. Their classification method utilizes locality-preserving projection, fuzzy c-means clustering, and soft-labeled support vector machines (SVMs). By employing complementary features on a larger codebook generated from smaller ones, they achieved improvements in both recall and precision for image retrieval.

Further advancements were made by ElAlami et al. [8], who developed a feature optimization technique combined with a genetic algorithm to address the high computational costs associated with large feature representations. This approach effectively eliminated irrelevant features during similarity assessments, thereby enhancing retrieval precision. Another innovative method by Chum et al. [6] introduced query image expansion along with a feature extraction technique known as the "bag of words," focusing on significant image regions. This method improved retrieval performance by concentrating on salient image features during holistic image analysis.

These advancements demonstrate a significant evolution from the basic QBIC method, incorporating sophisticated feature extraction and optimization techniques that have notably enhanced the efficiency and accuracy of CBIR systems. As research progresses, further integration with machine learning is expected to continue refining and strengthening image retrieval methodologies.

A modified query expansion approach for image retrieval, detailed in [5], utilizes a feature filtering correlation process to enhance the precision of search results, reducing ambiguity and improving retrieval accuracy. This method is particularly effective in narrowing down the scope of similar images retrieved from a database, addressing challenges associated with image ambiguity in CBIR systems. Another notable advancement is presented by Yang et al. [36], who developed an innovative mobile image retrieval scheme. This system employs Scale-Invariant Feature Transform (SIFT) features alongside a query expansion technique to retrieve similar images stored on a mobile device. The query object selection is based on metadata such as location and timestamps of saved images, making the process contextually relevant. However, while the reliance on key point extraction can increase retrieval precision, it also renders the system susceptible to errors when handling noisy images. Additionally, maintaining saliency during key point-based similarity assessment poses a significant challenge.

In another study, Garg et al. [10] introduced a CBIR method that integrates multi-level feature generation and compression to improve retrieval accuracy. Their approach involves multi-level image decomposition by extracting both approximation and detail coefficients through discrete wavelet transformation applied to individual color channels. The resulting structure is processed using local binary patterns, with the extracted magnitude providing additional discriminative power. The Gray-Level Co-occurrence Matrix (GLCM) method is then applied to these dominant patterns to generate statistical inputs for texture classification. Similar multi-level pipelines [25, 28] leverage Convolutional Neural Network (CNN) architectures to further enhance accuracy, particularly in specialized applications like medical image retrieval.

Aiswarya et al. [2] proposed an efficient CBIR model that leverages system memory to optimize the retrieval process. Their method addresses typical limitations of key point descriptor-based searches by incorporating Particle Swarm Optimization (PSO) with SIFT key points as feature descriptors. The system also uses query expansion to maintain image saliency, while dimensionality reduction techniques optimize the feature space by eliminating outliers. Despite these improvements, the model struggles with consistent performance across various image types. A key issue is that static feature extraction algorithms used in these models fail to deliver robust performance in all scenarios, emphasizing the need for a more dynamic approach that adapts to the specific characteristics of different image samples.

Managing high computation times, especially when working with large datasets on mobile devices with limited hardware capabilities, remains a concern. One way to address this is by limiting the number of reference images used during similarity searches, thus reducing computational overhead and making the process more feasible for resource-constrained environments. Another recent CBIR approach, presented in [1], incorporates an autoencoder for feature extraction and adaptive selection of relevant features from the target dataset. This model also uses query expansion to preserve visual saliency, achieving good retrieval accuracy. However, the processing time and computational cost remain high, indicating the need for further enhancements.

To address these limitations, this paper proposes a novel multi-level CBIR system. The proposed approach aims to overcome the challenges identified in previous methods by implementing a fast retrieval process based on features extracted from a deep learning image classification model. Specifically, the system employs clustering over the target space to reduce the number of reference images and then applies a CNN-based classification scheme to select the most relevant clusters for similarity assessment. This clustering mechanism not only optimizes the retrieval process by lowering computational demands but also enhances the precision of similarity searches by focusing on the most pertinent image subsets.

In the proposed scheme, clustering is performed on the feature space to group images into distinct categories, significantly reducing the search space and computational burden. The CNN model is trained on a diverse set of images, enabling it to effectively classify images into appropriate clusters based on their features. This method allows the CBIR system to quickly and accurately retrieve images that are contextually similar to the query image. By combining the strengths of both traditional and advanced CBIR techniques, this approach offers a comprehensive solution to the challenges of image retrieval.

A detailed explanation of the proposed algorithm and its implementation is provided in Section 3, where the multi-level strategy, the integration of clustering with CNN-based classification, and the performance improvements achieved through this hybrid approach are discussed in depth.

3. METHODOLOGY

The proposed image retrieval process involves categorizing input data into three datasets: global, local, and query image datasets. The global dataset consists of a large collection of real-world images sourced from various locations such as cloud storage, internet repositories, memory cards, and external hard drives. This dataset serves as the main pool from which the most relevant and similar images need to be extracted. The local dataset, on the other hand, includes images that are readily accessible within the immediate environment, such as those stored on a local machine or network. It allows for faster and more focused retrieval. If the local dataset is small, data augmentation techniques can be employed to increase its diversity and size, enhancing the training process. The query image acts as the primary input used to search for similar images within the master dataset.

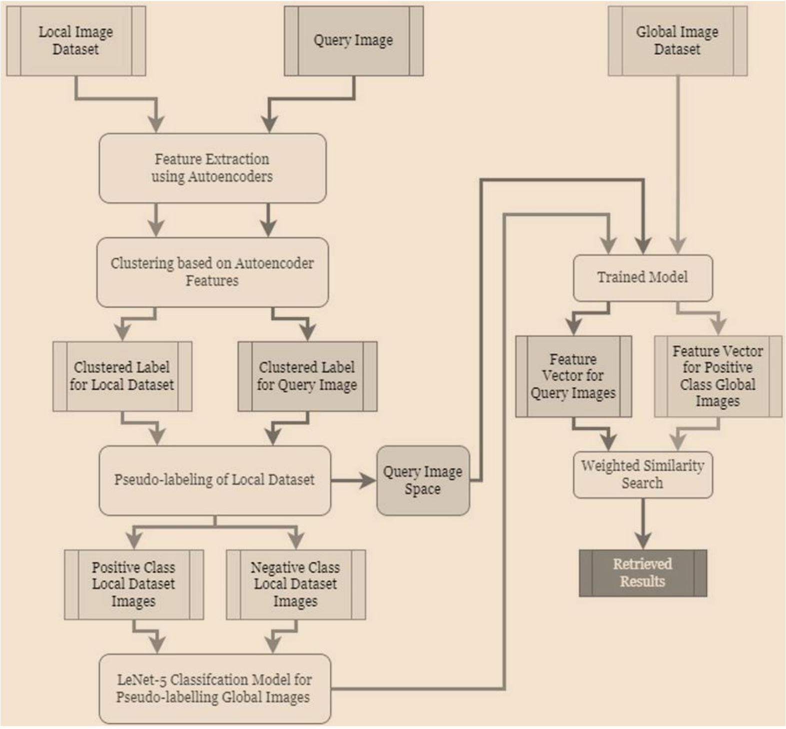

Local Image Dataset
Query Image
Global Image Dataset
Feature Extraction using Autoencoders
Clustering based on Autoencoder Features
Trained Model
Clustered Label for Local Dataset
Clustered Label for Query Image
Feature Vector for Query Images
Feature Vector for Positive Class Global Images
Pseudo-labeling of Local Dataset
Query Image Space
Weighted Similarity Search
Positive Class Local Dataset Images
Negative Class Local Dataset Images
Retrieved Results
LeNet-5 Classifcation Model for Pseudo-labelling Global Images

Fig. 3: Block diagram of the suggested CBIR system.

A key challenge in Content-Based Image Retrieval (CBIR) is the lack of labeled data, which is necessary for effective classification. Traditional machine learning models rely on labeled datasets to train the system to distinguish between different classes. However, CBIR systems typically lack such labeled data, making it difficult to utilize pre-trained image recognition models effectively. This limitation impedes the retrieval process, as there are no predefined labels to guide the search.

To overcome this challenge and improve image retrieval accuracy, the proposed methodology introduces a technique called pseudolabelling. Pseudolabelling involves assigning provisional labels—positive or negative—to images based on their similarity to the query image. This approach creates a pseudo-supervised learning environment, even in the absence of explicit labels, thereby enhancing the system's ability to distinguish relevant images.

The methodology is tested using the global, local, and query datasets, with pseudolabelling applied to the local dataset to establish a foundation for similarity assessment. To ensure accurate feature extraction during pseudolabelling, the method employs autoencoders. Autoencoders have demonstrated effectiveness in learning compact representations of input data and extracting precise features in previous experiments. In this process, the autoencoder-derived features of both the query image and the local dataset are used to label the local images as positive or negative, based on their similarity to the query image.

The CBIR system also utilizes an adaptive approach for feature extraction from the image pool. This flexibility allows the system to refine the feature extraction process according to the specific characteristics of the dataset, resulting in more accurate retrieval outcomes. A detailed block diagram of the proposed CBIR scheme is provided in Figure 3, and the methodology is further elaborated in subsequent sections, describing each stage of the process in a structured and comprehensive manner.

3.1 Initialize input data

A Content-Based Image Retrieval (CBIR) system fundamentally comprises two core components: the query image and the target database. The query image acts as the input, and the system retrieves similar images from the target dataset based on this reference. In the proposed multilevel aggregation approach, three distinct datasets are utilized: an expanded query set, a local image dataset (comprising images stored in the device's memory), and a global image dataset (encompassing a broader collection of images, whether online or offline).

ncorporating both local and global datasets proves particularly useful for retrieval systems on mobile devices. The local dataset consists of images that are directly accessible on the device or within connected storage, while the global dataset includes images stored on external sources like cloud storage or external hard drives. By leveraging data from both local and global repositories, the retrieval process becomes more precise, taking advantage of the diverse data available both on the device and from external sources.

To simulate the retrieval process, the entire dataset is divided into three distinct groups: test, local, and global. Ten percent of the dataset is set aside as the test set, which serves as the source of query images during evaluation. Thirty percent is designated as the local dataset, representing images that are immediately accessible on the device, while the remaining sixty percent makes up the global dataset, acting as the main search pool for retrieval. It is ensured that all three subsets have a balanced representation of the various image classes within the dataset.

Due to the absence of predefined ground truth information for the local, global, or query images, a pseudolabelling process is introduced. This step is intended to approximate the grouping of images within the input datasets based on their visual similarities. The primary goal of pseudolabelling is to identify additional query images from the local dataset, facilitating a more streamlined search process by narrowing down the number of target images considered during the final retrieval phase. This method not only speeds up the search process but also improves the accuracy of the results, especially when working with extensive and varied image collections.

3.2 Initial clustering

To create pseudolabels for the images in the local dataset, we utilized a k-means clustering approach based on features extracted from an autoencoder. An autoencoder is a specialized neural network that learns to represent data efficiently, often for dimensionality reduction or feature extraction purposes. It operates through an iterative process that establishes a low-dimensional latent space, which can reconstruct images with minimal reconstruction error. The autoencoder consists of two main parts: an encoder that converts input images into a feature vector and a decoder that reconstructs the original images from this feature vector. These functions are fine-tuned iteratively for a set of images, enhancing their performance over time.

In our approach, we leveraged the trained autoencoder to generate feature vectors for each image in the local dataset, as well as for the query image. If there are n images in the local dataset, the total number of images used in this process is $n + 1$ (including the query image). We selected a moderate feature vector size of 250 dimensions, providing a generalized yet effective representation of the image features.

Once we computed the feature vectors using the autoencoder, we moved on to the clustering operation. The main goal of this phase is to identify images in the local dataset that are similar to the query image. To accomplish this, we applied k-means clustering, organizing the images into multiple clusters based on their feature vectors.

After the clustering phase, we classified the feature vectors of the local images into two categories: positive and negative classes. The feature space for the local dataset, obtained from the autoencoder, consists of dimensions measuring $n * 250$. By adding the feature vector of the query image, we expand this to $(n + 1) * 250$. The clustering is executed with $k = 2$, resulting in cluster labels for each feature vector.

The output from the k-means clustering associates each $n + 1$ feature vector with a corresponding cluster label. The label assigned to the $(n + 1)$th image, representing the query image, is particularly important in the classification process. The cluster to which the query image belongs serves as a seed point for dividing the remaining images into positive and negative datasets. Images within the same cluster as the query image are classified as the positive class, while those in different clusters are categorized as the negative class. This classification establishes the first level of aggregation, with the positive and negative labels representing the pseudolabelling of the local dataset images. All subsequent stages of the retrieval process depend on these pseudolabels, laying the groundwork for more accurate image retrieval results. This method effectively enhances the retrieval system's ability to identify relevant images based on their similarity to the query input.

3.3 Creation of query image space

The query image space is an expanded collection of images created by adding a selection of similar images to the original query image. During this phase, we identify the top N images from the reduced positive class of local dataset images that are most similar to the initial query. A set similarity threshold is established to ensure that only images meeting this criterion are included in the query image space.

By augmenting the original search query with these selected local images, we form a robust query image space that complements the global repository in the content-based image retrieval (CBIR) framework. This method effectively generates a pseudo-labelled local dataset, enhancing the overall effectiveness of the retrieval system.

The final query set, represented as Q = $\{q_1, q_2, ..., q_N\}$, includes images ranked by their similarity to the original query image q_1. Here, q_1 is the actual query image, while q_2 is the most similar image from the local dataset, q_3 is the next most similar image, and so forth. The ranking of the images in the query space follows the order $q_1 > q_2 > q_3 > q_4 > ... > q_N$, facilitating a more precise and efficient search process for effective image retrieval.

3.4 Pseudo labelling of global dataset

In this phase, we assess the pseudolabels for each target image within the global dataset. The following are the detailed steps involved.

3.4.1. Training a CNN classification model

Since predefined labels are not available for the images in the global dataset, we constructed a deep learning classification model to make predictions for this dataset. This model is based on a Convolutional Neural Network (CNN) architecture, which is trained using local dataset images along with their generated pseudolabels. For our experiments, we utilized a customized version of the LeNet-5 convolutional neural network specifically designed for image classification.

To mitigate the challenge posed by the limited number of local images, we applied data augmentation techniques to sufficiently increase the dataset size, ensuring that the model is trained effectively without the risk of underfitting. The CNN architecture consists of three convolutional layers, each employing a kernel size of 5 × 5, along with two max-pooling layers and a fully convolutional layer. The first convolutional layer utilizes six filters, the second employs sixteen, and the final layer uses 120 filters to capture complex image features. All images are resized to 64×48 pixels with three color channels, striking a balance between accuracy and computational efficiency.

The model also incorporates batch normalization and dropout techniques to improve performance. The input layer processes images of size 64×48 with ReLU activation functions. The addition of max-pooling layers after the first and second convolutional layers effectively restricts the feature space. For optimization, the Adam optimizer is employed with a learning rate of 0.001, determined through empirical trials. The final classification layer uses SoftMax activation along with the categorical cross-entropy loss function to enable effective learning and classification outcomes.

3.4.2. Predicting global image labels from the trained model

After the CNN model has been trained, all images in the global dataset are processed through it for classification. Each image is categorized as either part of the positive class (those that are similar to the query image) or the negative class (those that differ from the query image). This pseudolabeling process effectively reduces the number of global images for further analysis. For instance, if the global dataset contains 1,000 images, with 400 identified as positive and 600 as negative, only the 400 positive images will be considered for the next steps. As a result, the final search space is confined to these positive global images along with the query image space established in the previous stages.

3.4.3. Generating feature vectors for final retrieval process

Once the positive class global images are identified, a refined set of feature vectors is produced from the output of the CNN's final layer. If each of the 400 selected images corresponds to 120 neurons in that layer, a global feature space of size 400 × 120 is created. This global feature space acts as the final search space for the retrieval process. Similarly, when processing up to 5 query images through the trained CNN model, a query feature space of size 5 × 120 is generated. This configuration allows for effective comparison and retrieval of similar images based on their extracted features.

3.5. Creating bags of query images

In this phase, the search query is evaluated against the positively mapped global repository. Rather than employing a straightforward image differentiation technique, a weighted similarity measure is utilized to compare the query feature space with the global feature space. This method enables the grouping of similar images in the global repository according to each feature vector from the query space. The comparison is conducted using similarity checks, with the Chi-square similarity measure chosen over the Root Mean Square (RMS) method due to its advantageous characteristics.

The images in the query space are compared to the reduced set from the global dataset based on an adaptively fixed threshold similarity that evolves as the algorithm runs. This process results in the creation of indexed bags, ordered by similarity scores. For instance, if there are q images in the query space, each possessing 120 feature vectors, the total size of the considered data amounts to q × 120 elements. This leads to the formation of q bags, each containing relevant search results.

Assuming *q* equals 5, the indices for the bags will range from Bag 1 (for the first query image, *q1* to Bag 5 (for the fifth query image, *q5*. Each query image is compared against the 400 global images, and the similar image indices are collected into their corresponding bags. For example, *q1* is compared to the global images, and the resulting similar image indices are stored in Bag 1, with the same process followed for *q2* to create Bag 2, and so on.

In an ideal scenario where all query images share similarities, the contents of all bags would reflect each other. However, in real-world applications, the items within each bag may differ since the images in the query space are not identical. To refine the results, the system identifies the most frequently occurring items across the bags, assigning weighted priorities. The weight for each bag is structured as follows: Bag 1 is given a weight of 1, Bag 2 receives a weight of 1/2, Bag 3 is assigned a weight of 1/3, Bag 4 is given a weight of 1/4, and Bag 5 is assigned a weight of 1/5.

For images appearing in all bags, their scores are calculated by summing the assigned weights, resulting in a total score of 1 + 1/2 + 1/3 + 1/4 + 1/5. Ultimately, a unique union of all images is compiled from the sets, and the priority of these images is determined based on their frequency across the bags, providing a refined approach to image retrieval.

Bags	**Score calculation**	**Final Score**
1	1+1/3+1/4+1/5	1.783
2	1+1/2+1/4	1.75
3	1+1/2+1/3	1.83
4	1/2+1/5	0.7
5	1/3+1/4+1/5	0.78

Table 1: Score calculation for setting the priority.

Let's take an example where Bag 1 contains global image indices 1, 2, and 3; Bag 2 includes indices 2, 3, and 4; Bag 3 has indices 1, 3, and 5; Bag 4 features indices 1, 2, and 5; and Bag 5 contains indices 1, 4, and 5. The unique union of these indices results in 1, 2, 3, 4, and 5. To compute the scores for each global image, we sum the weights based on their occurrences in the bags. For global image 1, the score is calculated as 1 + 1/3 + 1/4 + 1/5, which totals approximately 1.783. For image 2, the score is 1 + 1/2 + 1/4, giving a total of 1.75. The scores for images 3, 4, and 5 are similarly computed, yielding approximately 1.83, 0.7, and 0.78, respectively.

These scores are summarized in Table 1. Next, an adaptive similarity threshold is applied. If we set the threshold at 1, the proposed method will select the images corresponding to global image indices 1, 2, and 3. These selected images will be reported as the final retrieved results.

4. RESULTS & DISCUSSION

Both quantitative and qualitative assessments were carried out to demonstrate the effectiveness of the proposed model. The evaluation utilized the Oxford Buildings Dataset, which comprises nearly 5,000 color images organized into 17 distinct categories, each representing different landmarks in Oxford. For the experiments, we selected 50 images from 15 classes. Some sample images are shown in Fig. 4. To ensure an unbiased evaluation, we randomly constructed a query image set, local dataset, and global dataset by selecting 10%, 30%, and 60% of the total data, respectively. All experiments were executed using MATLAB® 2021A on a PC equipped with an Intel® Core i7-7700HQ CPU running at 2.80 GHz, 16 GB of RAM, and an NVIDIA® GTX 1050 graphics card with 4 GB of memory.

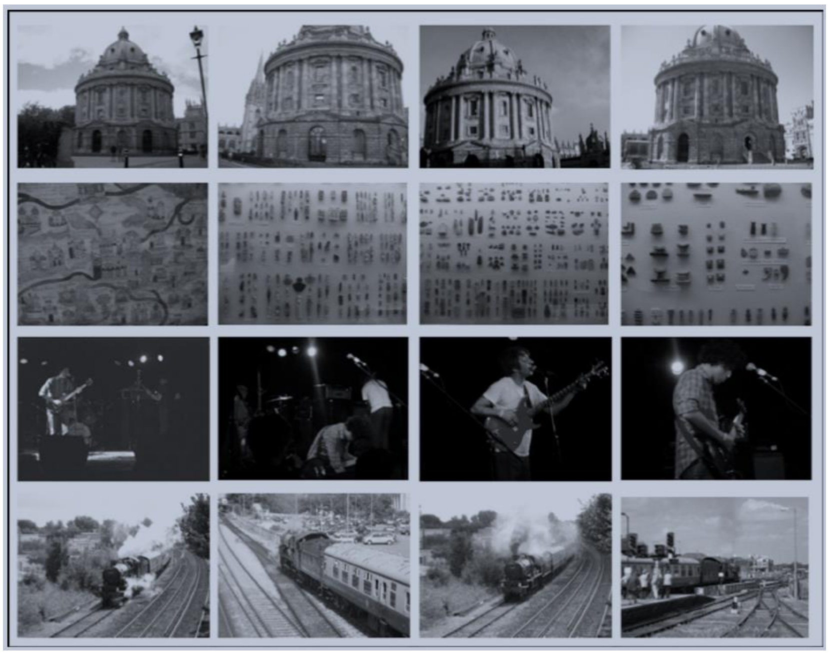


Fig. 4: A few sample images from the Oxford dataset that are used in the suggested CBIR system.

1. Average Precision (P_{avg})

Average precision gives the mean precision obtained while evaluating multiple random query images. If N_q is the number of query images then, the mean precision can be computed from the formula:

$$P_{avg} = \frac{1}{N_q}\sum_{i=1}^{N_q}\left(\frac{R_i}{N}\right)$$

where R_i represents the number of images from the *i-th* class that are retrieved among the top N retrievals.

2. Average True Positive Rate (TPR_{avg})

The Average True Positive Rate (TPRavg), also known as Recall, measures the average proportion of accurately retrieved samples from the total available data in the global dataset. The formula is as follows:

$$TPR_{avg} = \frac{1}{N_q}\sum_{i=1}^{N_q}\left(\frac{R_i}{min(N, M_i)}\right)$$

where *R_i* represents the number of correctly retrieved samples in the *i-th* class, and *M_i* denotes the total number of images in the *i-th* class among the top *N* results.

3. Average Error Rate (Err_{avg})

The average error rate reflects the mean error rate for all queries and is computed as:

$$Err_{avg} = \frac{1}{N_q}\sum_{i=1}^{N_q}\left(\frac{E_i}{N}\right)$$

where *E_i* represents the number of incorrect detections in the *i-th* class among the top *N* retrievals.

4. *Average False Positive Rate* (*FPR_{avg}*)

The False Positive Rate quantifies the average rate of incorrect retrievals from the false groups and is calculated as:

$$FPR_{avg} = \frac{1}{N_q}\sum_{i=1}^{N_q}\left(\frac{FP}{Neg_i}\right)$$

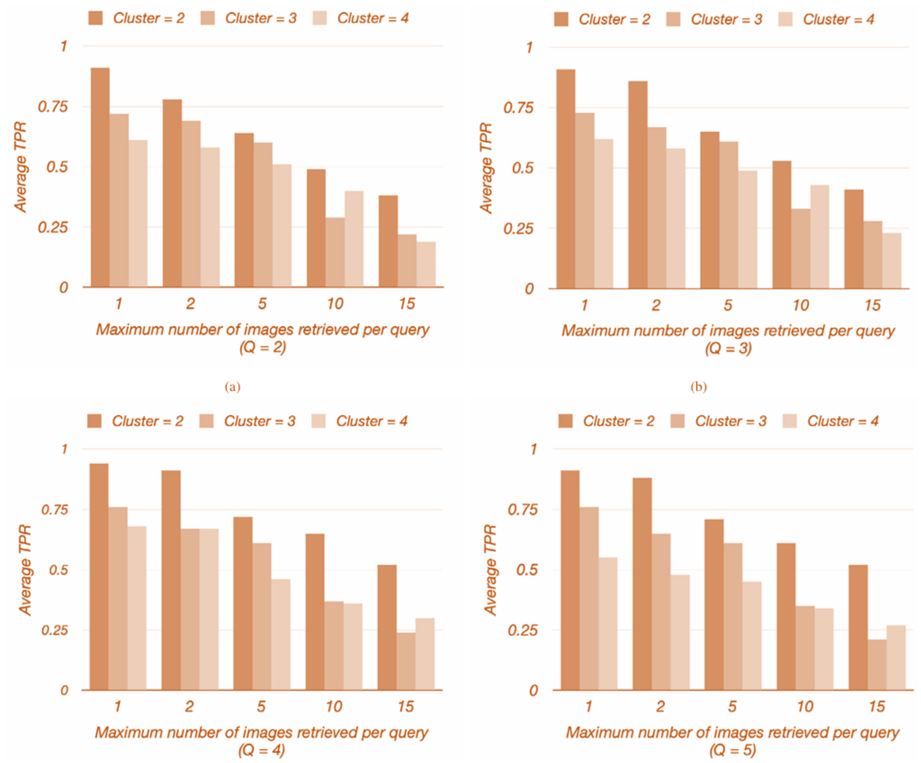
where *FP* denotes the number of false positives, and *Neg_i* represents the total number of negative samples in the *i-th* class.

The proposed method’s retrieval efficiency is significantly influenced by three key factors: the number of clusters used for pseudo-labeling, the feature vector utilized for evaluating image similarity, and the size of the query space. To understand their impact, several experiments were conducted by varying these parameters.

Figure 5 depicts how the Average True Positive Rate (TPR) varies with different numbers of clusters and query space sizes (Q). The highest average recall was observed when Q was set to 3, meaning the query space contained three images for the final comparison. Increasing the number of images in the query space generally led to a decrease in performance, highlighting that retrieving more images per query space negatively affects retrieval accuracy. Moreover, while increasing the number of clusters during pseudo-labeling enhances the likelihood of finding similar images in the local dataset, it also reduces the number of images in the positive class, thereby lowering recall rates. As a result, the number of clusters was treated as a hyperparameter, and through empirical testing, an optimal value of 2 was chosen, yielding reliable clustering outputs without excessively shrinking the positive class.

The initial number of clusters in the pseudo-labeling phase also has a significant impact on retrieval performance. The best results were generally achieved with fewer clusters, as increasing the number of cluster labels increases the risk of erroneously categorizing many positive images into the rejection class. This misclassification can degrade the final retrieval quality by excluding images that should have been included as positive matches.

Figure 6 provides an assessment of the average error rate under various conditions. The error rate remained consistently low (below 5%) when retrieving a smaller number of target images, regardless of the number of query images in the query space. However, the error rate began to rise when more than three query images were included in the query space. Additionally, retrieving a greater number of target images per query image led to a higher error rate, indicating a reduction in precision as the number of retrievals increased.



(a)

(b)

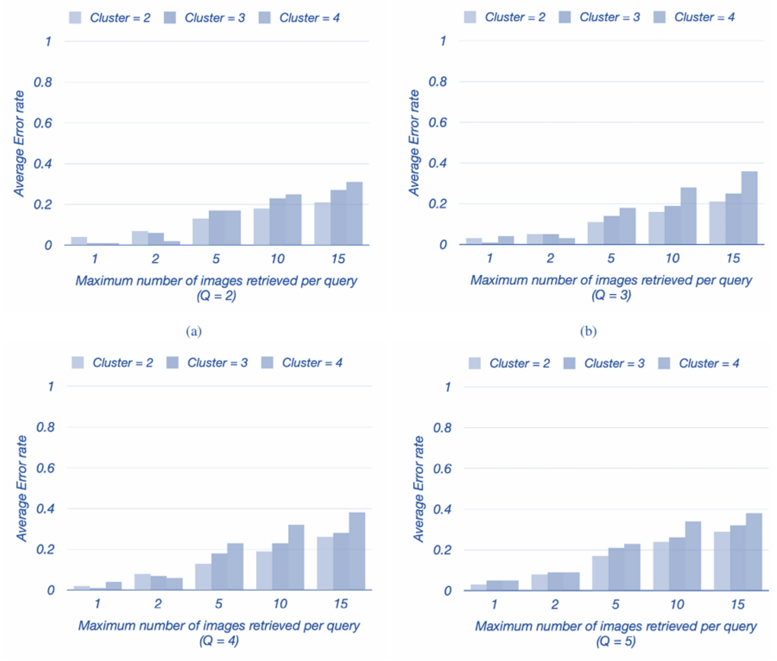
Figure 5: Average True Positive Rate with varying cluster counts, where Q denotes the total number of images in the query space.

To measure the classifier's effectiveness across different threshold levels, a Receiver Operating Characteristic (ROC) curve was employed, as shown in Figure 7. The ROC curve plots the True Positive Rate (TPR), or Sensitivity, against the False Positive Rate (FPR). The Area Under the Curve (AUC) quantifies the classifier's performance, with values ranging from 0 to 1, where 1 represents a perfect classifier. A higher AUC value indicates a more effective classifier in distinguishing between positive and negative classes. In this context, a high AUC value signifies the robustness of the proposed retrieval system.

Methods	Average Error Rate	Average TPR	Computation time (Sec)
[36]	0.57	0.53	5.9
[2]	0.48	0.61	7.11
Proposed	0.19	0.64	7.3

Table 2: Performance comparison of the proposed method with state-of-the art approaches (while retrieving 10 images per query image).

In the ROC curve experiments, an increase in the number of images within the query space initially led to an improvement in the Area Under the Curve (AUC), indicating better classifier performance. However, when the number of query images exceeded four, performance decreased due to a rise in false positives during the retrieval process. Table 2 provides a quantitative comparison of the proposed model with recent similar systems, highlighting its superior performance in terms of error rate and average True Positive Rate (TPR).



(a)

(b)

(c)

(d)

Figure 6: Average Error Rate with Various Cluster Numbers. Q denotes the total number of images in the query space.

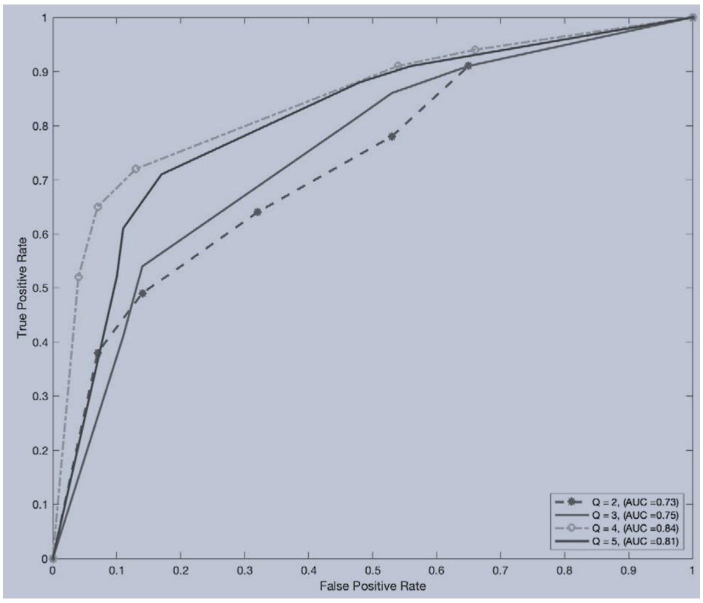


Figure 7: TPR versus FPRROC curve with two clusters for pseudolabelling, where Q is the number of images looked at in the query space.

Despite these benefits, the method requires slightly more execution time due to the high computational demands of certain stages. This trade-off between accuracy and computational efficiency is an important factor to consider in real-world applications.

5. CONCLUSION

The proposed Content-Based Image Retrieval (CBIR) model is designed to facilitate efficient image searches on electronic devices, including sources such as the internet and local storage. It utilizes a combination of advanced techniques to enhance retrieval performance while keeping computational demands low. By employing autoencoders to create image feature descriptors, the model effectively handles images with lower dimensions. The integration of k-means clustering with autoencoder-derived features allows for pseudo-labeling of images, reducing computational costs while enabling efficient feature analysis and categorization.

A key feature of the approach is its query expansion mechanism, which uses multiple query objects assigned with weighted priorities. This method leverages the available images in local memory to conduct a thorough exploration of the visual attributes of the query images. The combination of k-means clustering and query image expansion significantly reduces the likelihood of retrieving false samples, thereby boosting the overall accuracy of the results.

The deep learning-based image classification model further optimizes the process by accurately mapping query objects to relevant clusters. This targeted clustering restricts the search space, enhances retrieval precision, and reduces computational overhead. In the final phase, a weighted similarity assessment is performed on the query images, prioritizing query objects based on relevance. This helps minimize retrieval errors and ensures that the most pertinent images are selected.

Compared to traditional CBIR techniques, the proposed model exhibits superior performance across various quantitative metrics, including precision and recall. However, there is room for further improvement, particularly in boosting the recall rate. Future research could focus on integrating more advanced image labeling techniques to achieve higher recall without compromising average precision. Overall, the model represents a significant advancement in CBIR technology, providing a robust and efficient solution for image retrieval on electronic devices.

REFERENCES

[1] R. Bibi, Z. Mehmood, R.M. Yousaf, T. Saba, M. Sardaraz, A. Rehman, Query-by-visual-search: multimodal framework for content-based image retrieval, *Journal of Ambient Intelligence and Humanized Computing* 11 (2020) 5629–5648.

[2] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, in: *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, IEEE, 2005, pp. 886–893.

[3] Integration Of Renewable Energy Sources With Power Management Strategy For Effective Bidirectional Vehicle To Grid Power Transfer, Pollepale Siddhartha, Thokala Sujeeth, Bandari Shiva, J. Ramprabhakar.

[4] Classification of Breast Thermal Images into Healthy/Cancer Group Using Pre-Trained Deep Learning Schemes, Seifedine Kadry, Rubén González Crespo, Enrique Herrera-Viedma, Sujatha Krishnamoorthy, Venkatesan Rajinikanth.

[5] K. T. Ahmed, S. A. H. Naqvi, A. Rehman, and T. Saba. 2019. Convolution, approximation and spatial information based object and color signatures for content based image retrieval. In Proceedings of the International Conference on Computer and Information Sciences (ICCIS'19). 1–6. https://doi.org/10.1109/ICCISci.2019.8716437

[6] S. Aksoy and R. M. Haralick. 2000. Probabilistic vs. geometric similarity measures for image retrieval. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'00), Vol. 2. 357–362. https://doi.org/10.1109/CVPR.2000.854847

[7] L. Han *et al.* Multi-view local discrimination and canonical correlation analysis for image classification Neurocomputing (2018).

[8] W. Liu *et al.* Multiview dimension reduction via hessian multiset canonical correlations Inf. Fusion (2018).

[9] R. Zhu, X. Li, X. Zhang, M. Ma, MRI and CT medical image fusion based on synchronized-anisotropic diffusion model, *IEEE Access* 8 (2020) 91336–91350.

[10] Caron, M., et al.: Emerging properties in self-supervised vision transformers. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 9630–9640 (2021)

[11] S. Cui, L. Mao, J. Jiang, C. Liu, S. Xiong, Automatic semantic segmentation of brain gliomas from MRI images using a deep cascaded neural network, *J. Healthcare Eng.* 2018 (2018) 1 14Mar.

[12] Roman Bachmann, David Mizrahi, Andrei Atanov and Amir Zamir, *MultiMAE: Multi-modal multi-task masked autoencoders*, 2022.

[13] Hassan Akbari, Liangzhe Yuan, Rui Qian, Wei-Hong Chuang, Shih-Fu Chang, Yin Cui, et al., *Vatt: Transformers for multimodal self-supervised learning from raw video audio and text*, 2021.

[14] Chen, M., et al.: Generative pretraining from pixels. In: Proceedings of the 37th International Conference on Machine Learning, pp. 1691–1703. PMLR (2020). iSSN: 2640–3498

[15] G. Sumbul and B. Demir, "Plasticity-stability preserving multi-task learning for remote sensing image retrieval," IEEE Transactions on Geoscience and Remote Sensing, vol. 60, no. 5620116, pp. 1–16, 2022.

[16] Z. Xie, Z. Zhang, Y. Cao, Y. Lin, J. Bao, Z. Yao, Q. Dai, and H. Hu, "Simmim: a simple framework for masked image modeling," in IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 9643–9653.