Computer Vision and Pattern Recognition

(EEE3032)

Visual Search of an Image Collection

Rajnish Kumar

Master of Science in Artificial Intelligence

from the

University of Surrey

School of Computer Science and Electronic Engineering

Faculty of Engineering and Physical Sciences

University of Surrey

Guildford, Surrey, GU2 7XH, UK

November 2024

Supervised by: Prof Miroslaw Bober

Table of contents and Abstract

Contents

Abstract

This study assesses visual search techniques by evaluating descriptors, distance metrics, and detectors, identifying strengths and limitations for each. The global color histogram descriptor exhibits limitations in distinguishing overlapping color categories, whereas the spatial grid histogram, combined with PCA and Mahalanobis distance, improves performance through enhanced feature separation and texture handling. The SIFT detector achieves high precision for visually distinct categories but encounters misclassification challenges with overlapping categories. Overall, the spatial grid histogram with PCA and Mahalanobis distance proves to be an effective approach for image retrieval in datasets with mixed color and texture features, addressing challenges in image retrival and good precision more effectively than other descriptors tested.

Description of Visual search techniques implemented

1. Global colour histogram

The global color descriptor captures an image's overall color profile by computing a histogram across the RGB channels, treating the entire image as a single unit without considering spatial details. Each color channel is divided into bins, often using an 8x8x8 quantization, balancing detail and efficiency. This descriptor identifies images with similar overall color patterns, such as distinguishing blue-sky landscapes from those with neutral tones. Euclidean distance measures similarity between color distributions in query and candidate images.

Efficient and effective for datasets where color is a primary differentiator, this descriptor works well for large collections. Its simplicity makes it ideal for images with distinct, dominant colors, like landscapes. However, it lacks spatial information, meaning it may retrieve visually different images that share similar colors. This reduces its accuracy for complex images with multiple objects, and it is sensitive to lighting variations, which can impact color consistency across images.

2. Spatial Grid (Colour and Texture)

The spatial grid approach for color and texture is a feature extraction technique that divides an image into smaller cells, allowing feature analysis in specific regions rather than the entire image. By creating a grid, the method captures localized color and texture features from each cell independently, and these are combined to represent the image. Color features are often extracted using histograms within each cell, creating a detailed profile that surpasses a global histogram. For texture, angular quantization levels can be applied to detect patterns and orientations within each region, enhancing the ability to recognize intricate textures across the image.

This method's main advantage is its ability to capture spatial variations, making it effective for analyzing complex scenes with varied objects and textures. However, extracting features from multiple cells can be computationally demanding, especially with finer grids, and focusing on local features may sometimes overlook important global context. Balancing grid size and feature complexity is key to maximizing effectiveness without overloading computational resources.

3. Principal Component Analysis (PCA)

Principal Component Analysis (PCA) is a dimensionality reduction technique that transforms high-dimensional data into a smaller set of components, preserving essential patterns while reducing complexity. By identifying principal components directions of greatest variance PCA projects data along these axes, eliminating redundant features and filtering out noise, making the data easier to analyze. PCA's primary advantage is that it enables efficient data processing, retaining key information while reducing the number of features. This makes it highly effective for large datasets, such as in image recognition and gene analysis, where reduced dimensions can speed up computations and improve performance. PCA also simplifies complex data, making patterns easier to visualize.

However, PCA has limitations. Reducing dimensions can lead to information loss, especially if too many components are removed. Since PCA is a linear method, it is best suited to linearly structured data, and finding the optimal number of components is crucial to maintaining data integrity.

4. Bag of Visual Words (BoVW)

The Bag of Visual Words (BoVW) retrieval system is used in image recognition by representing images as collections of visual words or features, similar to words in a document. The process begins with detecting keypoints in an image using feature detectors like SIFT, which identify distinctive points for comparison across images. These keypoints are then described by feature vectors that capture local appearance. To organize descriptors, k-means clustering groups similar ones into clusters, forming a codebook where each cluster center represents a visual word. The image is then represented as a histogram of these visual words, indicating their frequency. BoVW is effective for image retrieval, comparing images based on their visual word histograms. Its advantages include simplicity, robustness to transformations like scale and rotation, and ability to capture global image structure. However, it loses spatial relationships between features, which can make it less effective at distinguishing images with similar visual words but different spatial layouts.

Distance Metrics

1. Manhattan Distance

The L1 norm, also known as Manhattan distance, measures the distance between two points by summing the absolute differences of their coordinates. It is especially useful for high-dimensional and sparse data (lot of missing values or zero values), where individual dimensions are important.

One advantage of this metric is that it is less sensitive to outliers compared to Euclidean distance, making it ideal for sparse data. However, it may not capture the true geometric distance, especially in continuous spaces, and can be less accurate in cases where a straight-line distance is more meaningful.

2. Euclidean distance

The L2 norm, or Euclidean distance, calculates the straight-line distance between two points and is the most commonly used metric in continuous spaces. Euclidean distance is intuitive and works well when differences in each dimension should have a direct impact on the distance. However, it is sensitive to outliers, as large differences in any dimension disproportionately increase the total distance. Additionally, in high-dimensional spaces, its effectiveness diminishes where the data becomes sparse and less distinguishable.

3. Mahalanobis distance

The Mahalanobis distance measures the distance between a point and a distribution, accounting for correlations between variables. It is particularly useful for identifying outliers and analyzing multivariate data.

This metric is advantageous because it normalizes the distances based on the data's covariance, making it effective for correlated variables. However, it requires calculating the covariance matrix, which can be computationally expensive, particularly in high-dimensional settings (that is why it is recommended to use PCA before calculating this distance).

4. Cosine

Cosine similarity, which focuses on the angle between feature vectors rather than their actual distance, works best when the direction of features holds more importance than their size. However, for high-dimensional, PCA-transformed image descriptors, it doesn't perform as well. This metric often lacks the precision needed for clear feature separation, leading to lower accuracy in both retrieval and classification compared to metrics like Euclidean or Mahalanobis distance. These other metrics handle complex datasets more effectively by providing better differentiation between similar and distinct images.

Evaluation

1. PR curve

In the context of image retrieval, a Precision-Recall (PR) Curve is used to evaluate the performance of a system when retrieving images based on a query image. It plots the trade-off between precision and recall as more images are retrieved. Precision measures how many of the retrieved images are relevant, while recall measures how many of the relevant images in the dataset have been retrieved. As more images are retrieved, precision typically decreases, while recall increases.

The PR curve helps assess the effectiveness of an image retrieval system by showing how well it retrieves relevant images at different retrieval ranks. The area under the PR curve or Average Precision (AP) is often used as a summary metric, representing the system's overall retrieval performance. The PR curve is particularly useful when dealing with imbalanced datasets, where relevance to the query is more important than the total number of retrieved images.

2. Confusion Matrix

The confusion matrix was used to evaluate the accuracy of the retrieval system by comparing the predicted categories of retrieved images with the true categories. It provides an insight into how well the system distinguishes between different categories. The diagonal elements represent correct predictions, where the retrieved images match the true categories. Off-diagonal elements indicate misclassifications, where the retrieved images belong to categories different from the true ones. This matrix helps identify specific categories where the system performs well and others where it may need improvement.

Experimental results

1. Global colour histogram

Implementation details

This code implements a visual search system for image retrieval using the global color histogram technique. It processes a dataset of images by extracting their color histograms as descriptors, which are saved as .mat files. For each query image, the distance between its color histogram and those of all other images in the dataset is computed, with the top 10 retrieval results displayed. Precision and recall values are calculated for each query image, and precision-recall curves are generated to evaluate the system's performance in retrieving relevant images. A confusion matrix is also generated to compare the true categories of the query images with the predicted categories of the top 10 retrieved images, providing a measure of retrieval accuracy. These results and the confusion matrix are visualized using plots and color maps.

In this study, five selected query images are used to assess the retrieval performance of the global color histogram approach. These query images are chosen to represent different categories, and the top 10 matching results are retrieved for each. The precision-recall (PR) curves for each query image, plotted against the entire dataset, demonstrate the model's ability to distinguish relevant images from irrelevant ones. This provides an overall evaluation of the system's retrieval performance across various categories. Additionally, the confusion matrix helps assess the accuracy of category-level predictions. Since time constraints prevent the inclusion of subcategories (e.g., grass, cow, dog), the analysis focuses only on the primary 20 categories. This ensures a coherent and focused evaluation of the retrieval results based on these main categories. To maintain consistency, this study uses five query images, along with the PR curve and confusion matrix, as standard for comparison across all descriptors, including the global color histogram.

Quantization

Adjusting the color quantization level significantly affected the global color descriptor's performance. At lower levels (e.g., 4), the descriptor struggled with categories like 1 (grass, cow) and 4 (aeroplane, grass), as it focused on background colors instead of objects. As quantization increased to 8 and 16, retrieval accuracy improved, particularly for categories with diverse colors, like Category 13 (books). Performance peaked at quantization level 32, capturing finer details without excessive noise. However, higher quantization levels introduced noise, reducing accuracy. Thus, quantization level 32 was chosen for optimal results.

Results

The global color histogram performs poorly for Category 1 (grass and cow) because grass appears in multiple categories, leading the model to focus more on the background than the objects. This results in difficulty retrieving the cow, as grass dominates the image. The model prioritizes grass, causing a significant drop in precision in the PR curve for this category (Figure 2). This illustrates the descriptor's limitation when objects share similar colors or are overshadowed by background elements. Similarly, in Category 2 (aeroplane), the presence of white objects like buildings and signboards confuses the model. For Category 6 (figure 1) this descriptor is able to retrieve some good results but Despite this, the PR curve (Figure 2) shows poor results due to overlapping features between the two categories (category 6 and 19 are same human). In contrast, for Category 8 (bike and building) and Category 13 (book), the descriptor performs well, retrieving relevant images with high precision. Occasionally, books from other categories are included due to similar visual features, but the results are still generally good.

The global color histogram excels when distinct, large colors (e.g., sky, grass, ocean) dominate the image. However, it struggles when objects have similar colors or there are multiple objects in an image

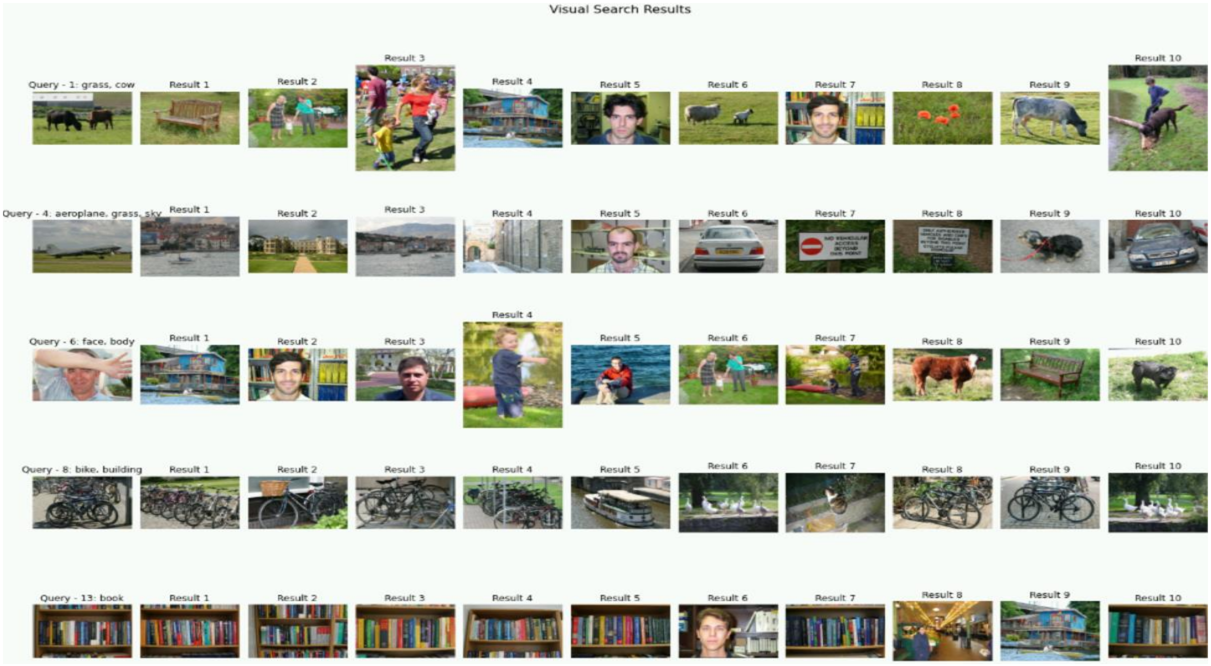

Figure 1 : Visual results of the five query images along with their top 10 retrieved images based on the global color histogram.

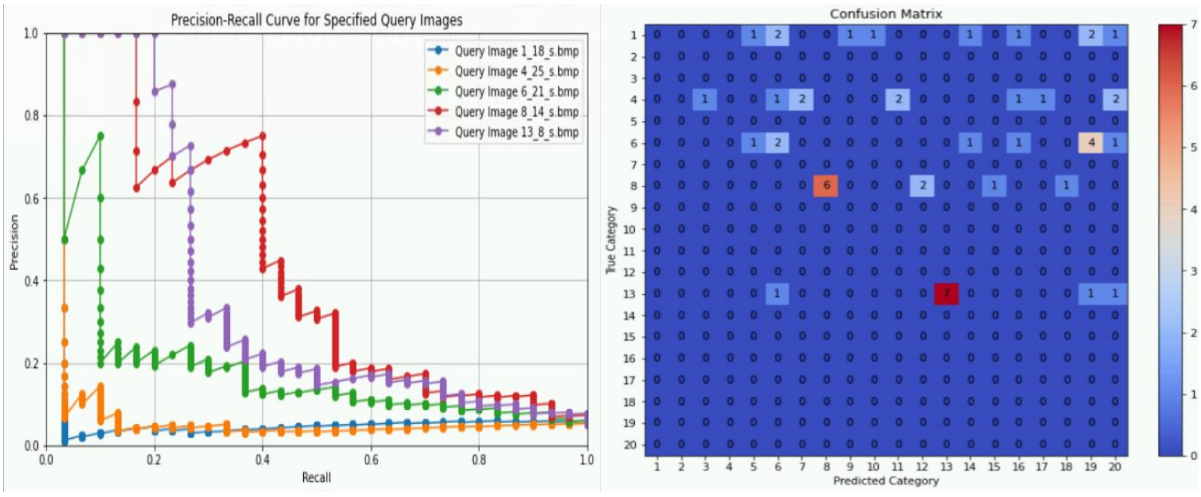

Figure 2 : Precision-Recall curve and confusion matrix for the global color histogram descriptor, showing the retrieval performance across 5 query images.

Overall, the descriptor is effective in categories with clear, distinct colors but struggles with complex images or those containing multiple objects with similar colors. It is fast, simple, and works particularly well for categories like Categories 8 and 13. However, its performance for grass-related categories remains low, potentially due to grass appearing in many categories, which leads to reduced retrieval accuracy.

2. Spatial Grid (Colour and Texture)

The Spatial Grid (Color and Texture) descriptor is designed to capture local features of an image by dividing it into grid cells, allowing more nuanced image representation. Through varying grid sizes, color bins, and angular quantization levels, the descriptor's performance can be fine-tuned for different image characteristics. Implementation details are somewhat similar to global color histogram. Below is a summary of findings based on the experiments conducted.

Grid Size

Increasing grid size generally improved performance, with results peaking around a 4x4 grid. Beyond this, improvements became minor, and higher grid sizes such as 16x16 and 32x8 did not lead to significant gains, with the 64-bin grid size consuming excessive memory.

Color Bins

At 4-color and 8-texture, categories 1 (grass, cow) and 6 (human face) had lower PR curves, while categories 4, 8, and 13 showed moderate success. Increasing to 8-color and 8-texture improved performance overall, though challenges persisted for categories 1 and 6. Small improvements were noted with 16-color and 8-texture, but gains diminished. At 32-color and 8-texture, improvement was minimal, with high memory usage, showing that further increases in color levels had limited effect on accuracy. These results suggest that while increasing color values initially boosts accuracy, higher values beyond a certain point do not contribute substantially to performance and may not be efficient for resource usage.

Angular Quantization (Texture)

For texture settings, 16 Color, 1 Texture showed poor results across most categories, with lower PR curve values, highlighting the significant role of texture in improving accuracy. 16 Color, 4 Texture led to moderate improvement, particularly for the aeroplane category, though categories 1 and 6 still struggled. 16 Color, 8 Texture saw an upward trend in PR curves, with categories 8, 13, and 4 benefiting the most. With 16 Color, 16 Texture, there was a small improvement, and the confusion matrix showed clearer patterns of error. However, at 16-texture and higher, a significant amount of memory was needed to store feature data, and there were minimal improvements in results beyond this level.

Results

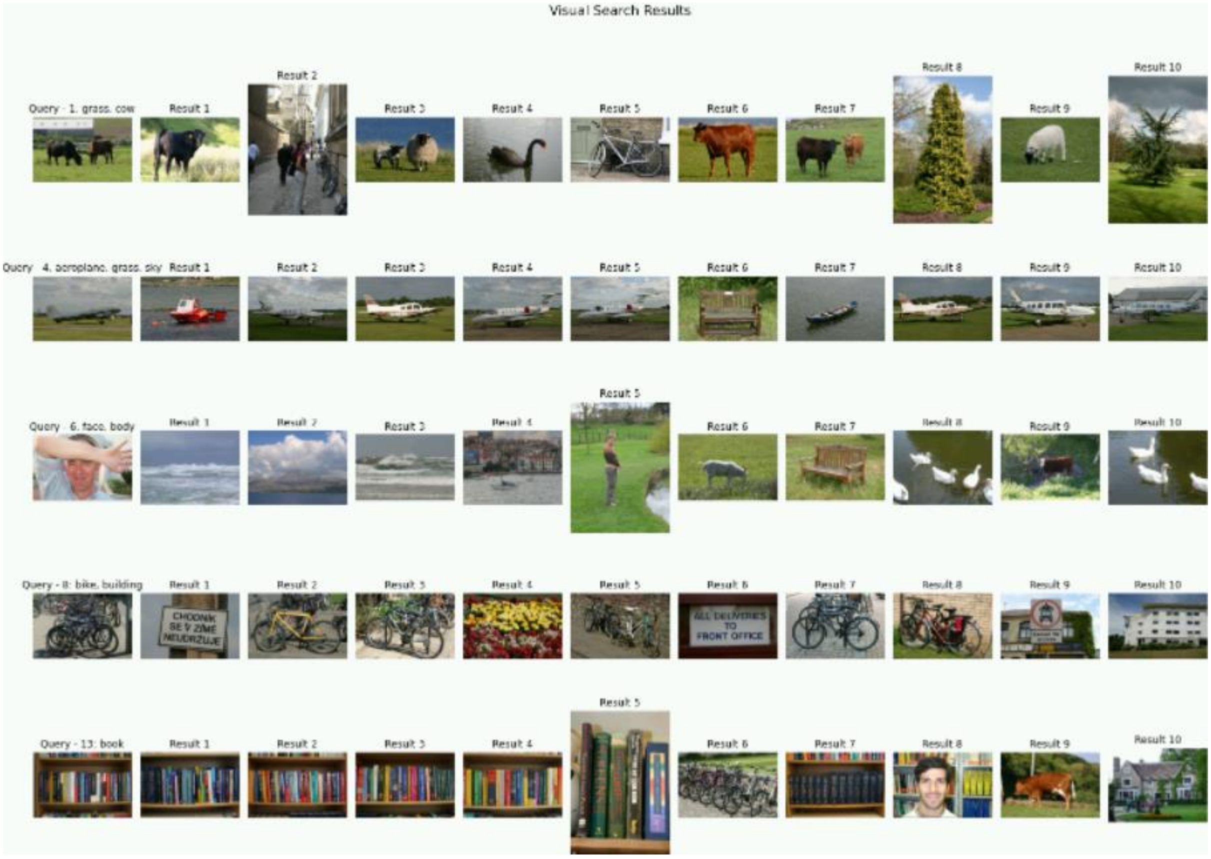

Figure 3 : Visual results of the five query images along with their top 10 retrieved images based on the Spatial Grid.

Comparing the results across the five primary categories, the model shows improved performance over the global color histogram approach due to its focus on localized color and texture features. However, Categories 1 (grass and cow) and 6 (face and body) still show lower performance, likely because grass and human elements frequently appear in the backgrounds of multiple categories (Figure 3). Although the model can localize objects, it often fails to differentiate them accurately. Additionally, overlapping subcategories lead to confusion in the model, particularly with animals, as shown in the confusion matrix (Figure 4), where animal types are often misclassified.

Distinct categories such as books, aeroplanes, and cycles perform well. For these, the PR curve generally centers in the mid-range, and the confusion matrix indicates that most of the top 10 retrieved results are correctly classified within the same category, demonstrating strong consistency. Notably, Category 13 (books) achieves high accuracy, with most top 10 results accurately categorized. For Category 8 (cycle and building), some misclassifications occur due to shared background elements like buildings and signboards, but overall retrieval remains consistent.

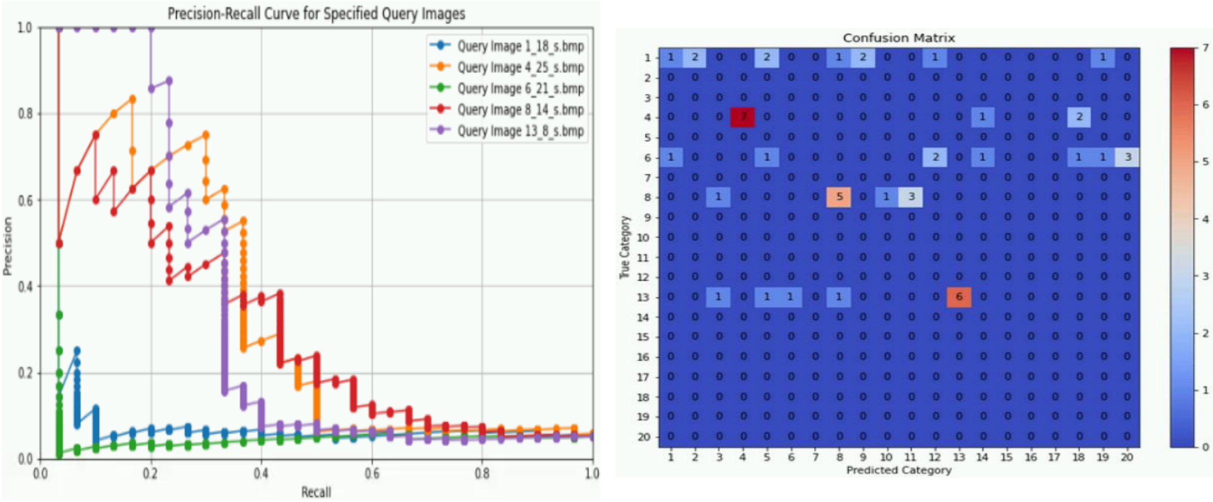

Figure 4 : Precision-Recall curve and confusion matrix for the Spatial Grid descriptor, showing the retrieval performance across 5 query images

The PR curve and confusion matrix reveal where the model excels, particularly in distinct categories, and where overlapping features reduce accuracy. Overall, the model outperforms the global color histogram approach in distinguishing unique categories but still struggles to differentiate animals or objects with similar backgrounds. This focus on background features helps for unique classes but limits retrieval accuracy when background and object features overlap significantly.

3. PCA

Implementation

This implementation focuses on building a visual search system by leveraging PCA for dimensionality reduction of image features. After extracting and loading image descriptors (Spatial Grid), PCA is applied to reduce the dimensionality, which helps improve computation efficiency, particularly when calculating Mahalanobis distance. The system then calculates distances between a query image and all others, retrieving the top 10 most similar images based on the chosen metric (e.g., Euclidean, Manhattan, Cosine, or Mahalanobis). Precision and recall are computed for each query, and the results are plotted in precision-recall curves to evaluate performance. Additionally, a confusion matrix is generated to analyze the accuracy of predicted categories, comparing the true and predicted labels for the top retrieved images. The overall setup enables experimentation with different distance metrics and provides insight into the effectiveness of PCA in enhancing search accuracy and computational efficiency.

PCA components

As the number of PCA components increases, the performance of the model improves up to a point, after which it stabilizes. With only 2 components, the precision-recall curves are quite low, reflecting a high level of misclassification due to significant information loss. The results at 5 components show little change, maintaining a similar level of performance. A modest improvement is observed at 10 components, particularly for the aeroplane category, though misclassifications are still prevalent. At 20 components, the model's performance improves noticeably, especially for the "book" category, where there are more true positives. This trend continues with 30 components, where categories 4, 18, and 13 see a marked improvement in classification. With 50 components, performance continues to improve, but the gains become less significant. Finally, at 70 components, there is no further improvement, and the performance stabilizes.

Results

For the results, this study used 50 components as the standard for PCA. The outcomes were similar to those observed with the Spatial Grid (Euclidean distance). The precision-recall curves indicate that categories 4, 8, and 13 performed better, as their curves were more centered, suggesting that images from these categories were retrieved more accurately compared to categories 1 and 6. By reducing the dimensionality of the spatial grid descriptor features, PCA maintained similar results while simplifying the data representation. This suggests that PCA was effective in retaining the most important information while reducing unnecessary complexity. Additionally, PCA proved especially beneficial when Mahalanobis distance was used, as it significantly reduced the computational cost. Without PCA, calculating the covariance matrices for Mahalanobis distance was taking too much time. PCA, by reducing the dimensionality, minimized both the computational burden and the noise in the data, leading to a more efficient and accurate model.

4. Distance measure

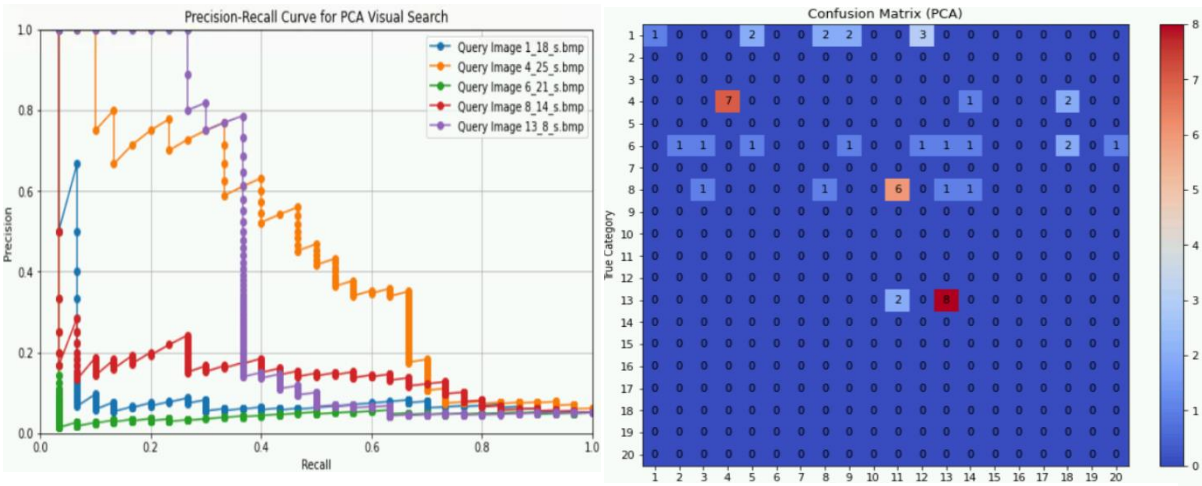

Figure 5 : Precision-Recall curve and confusion matrix after applying PCA with Mahalanobis Distance, illustrating retrieval performance across five query images.

Mahalanobis distance

The descriptor's performance shows only slight changes when switching from Euclidean to Mahalanobis distance, though varying the number of PCA components (n_components) reveals distinct patterns for each metric (Figure 5). With Mahalanobis distance, performance initially improves as n_components increases, reaching a peak around 18 components before declining. This suggests that Mahalanobis distance achieves optimal feature separation with fewer components, likely due to its sensitivity to the covariance structure, which can degrade as dimensionality increases.

Conversely, with Euclidean distance, performance continues to improve with more components, peaking around 50 components. This indicates that Euclidean distance benefits from higher-dimensional representations, as it does not rely on covariance and thus can capture more information without degradation.

In terms of precision-recall (PR) curves, Categories 13 and 4 show improvement with the new metrics, while Category 8 exhibits a decline. The confusion matrix reflects slight improvements in classification for certain categories, such as Category 13, with minor reductions in misclassifications. Overall, however, the improvement is modest, indicating that changing the distance metric has limited impact on the model's performance.

Manhattan

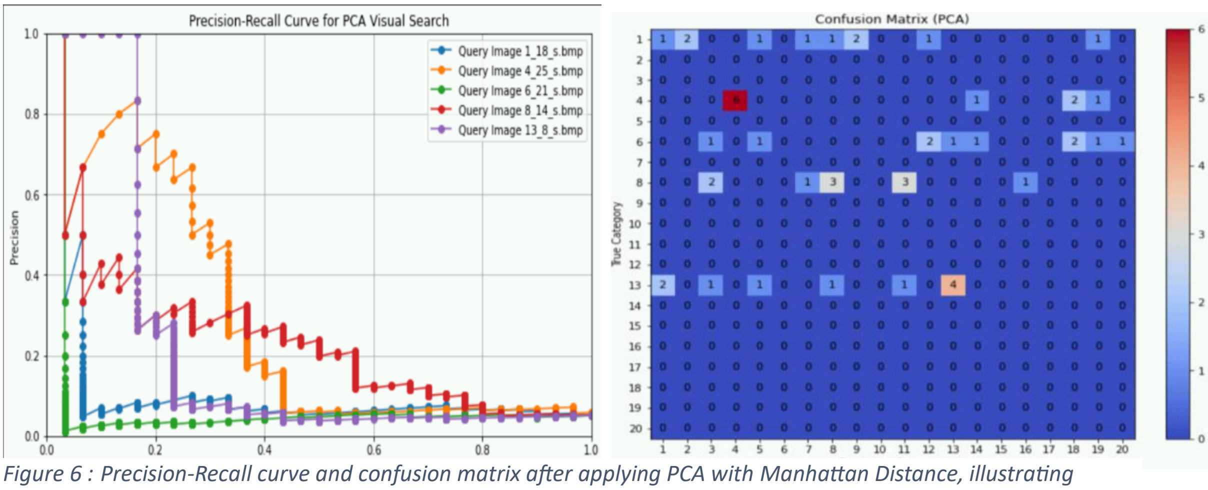

Figure 6 : Precision-Recall curve and confusion matrix after applying PCA with Manhattan Distance, illustrating retrieval performance across five query images.

Using Manhattan distance, the descriptor's performance declines noticeably compared to Euclidean distance. The precision-recall (PR) curve indicates that relevant images are not retrieved as accurately, showing reduced precision and recall for each query category. Additionally, the confusion matrix reflects an increase in misclassifications across categories, with each selected query category displaying a noticeable drop in retrieval performance.

The decline in performance with Manhattan distance is due to its nature. Unlike Euclidean distance, which measures straight-line distance, Manhattan distance uses the sum of absolute differences, which can be less effective in high-dimensional data. PCA-transformed descriptors work better with Euclidean or Mahalanobis metrics, which capture variance and correlations between dimensions. Manhattan distance oversimplifies these relationships, resulting in weaker feature separation and poorer retrieval and classification

Cosine

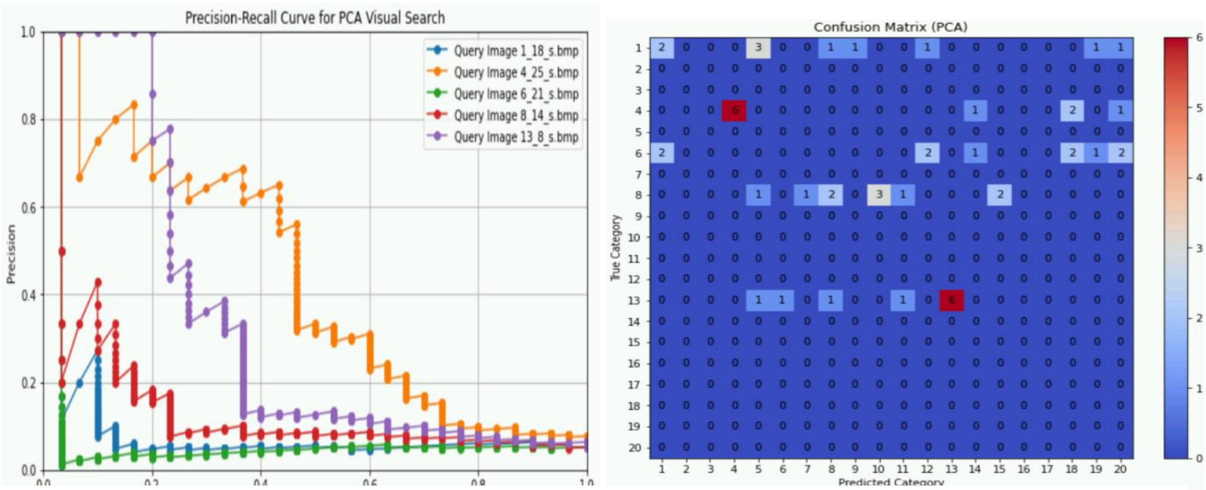

Figure 7 : Precision-Recall curve and confusion matrix after applying PCA with Cosine Distance, illustrating retrieval performance across five query images.

Compared to Euclidean distance, both the precision-recall (PR) curve and the confusion matrix show a decline in performance with cosine similarity. This decrease occurs because cosine similarity measures the angle between vectors rather than their absolute distance, which can be less effective for distinguishing images when the dataset includes high-dimensional PCA-transformed descriptors. As a result, cosine similarity may not capture variations in magnitude as well as Euclidean distance, leading to reduced retrieval accuracy and more misclassifications in the confusion matrix.

5. SIFT Detector

Implementation

This script performs a visual search of image collection using SIFT features. It extracts key point descriptors from each image, clusters them using K-means to create a codebook of visual words, and generates histograms that represent the frequency of these visual words in each image. For 5 query images, the histograms are compared to all others in the dataset using Euclidean distance to rank the images by similarity. The top 10 results are displayed alongside the query, and the performance is evaluated using precision-recall curves. A confusion matrix is also generated to assess the classification accuracy across 20 categories top 10 results, showing how often images from each category are correctly or incorrectly classified. The overall goal is to create a system that efficiently retrieves similar images based on visual content.

N_features

Increasing the number of n_features initially improves the performance by capturing more key points from the images, but this improvement plateaus after a certain threshold. For instance, at 1000 features, the performance improves significantly, but beyond this point, further increases in features offer little to no gain and might even cause some degradation. This suggests that there is an optimal number of features that balances the amount of information captured without overloading the model and computation cost.

Clusters

As for the number of clusters, increasing NUM_CLUSTERS refines the visual vocabulary used to classify the images. At smaller cluster sizes, such as 10 or 30, some categories, like 1 and 6, show lower performance, indicating that too few clusters may not capture enough variation. However, as the number of clusters increases to 70, 90 and 110, overall performance improves, and most categories are classified correctly. A further increase to 130 clusters leads to diminishing returns, and performance starts to decrease, especially when the number of clusters becomes too large. This suggests that while more clusters provide finer granularity, there is a point at which the increased complexity leads to overfitting. Cluster 110 is selected for results.

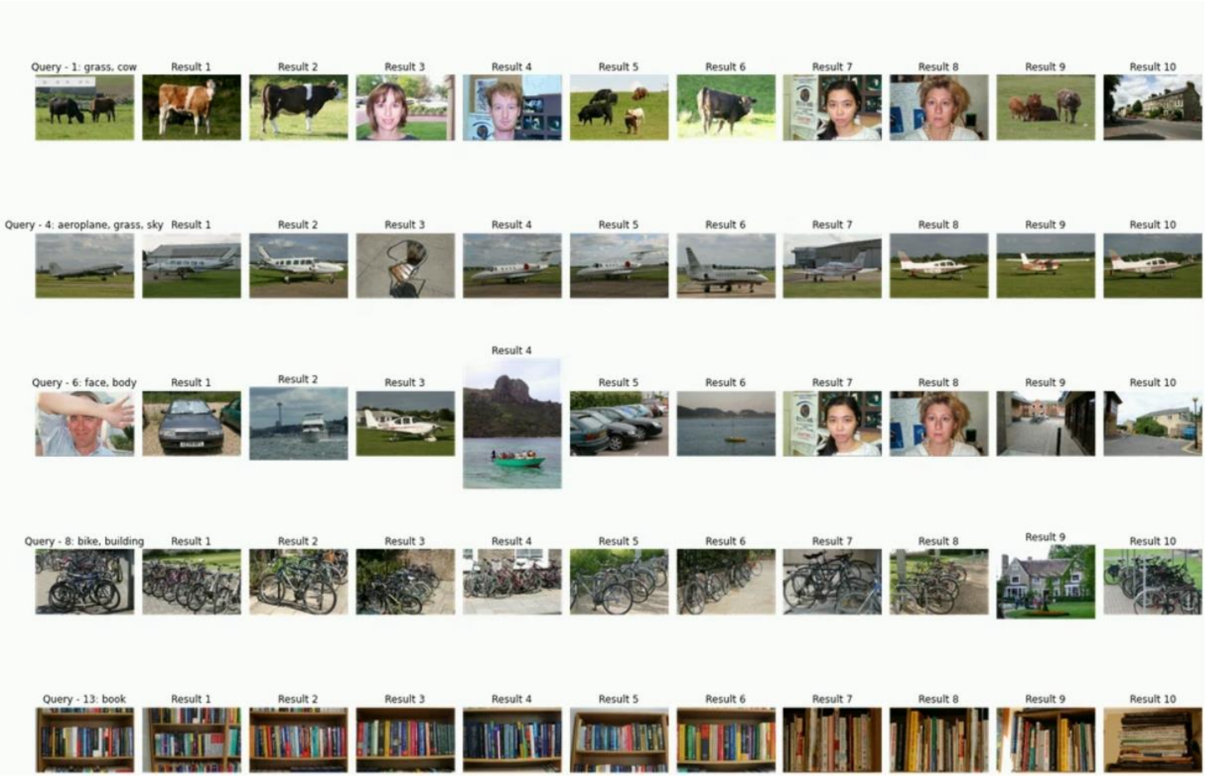

Figure 8 : Visual results of the five query images along with their top 10 retrieved images based on the SIFT Detector (Bag of words)

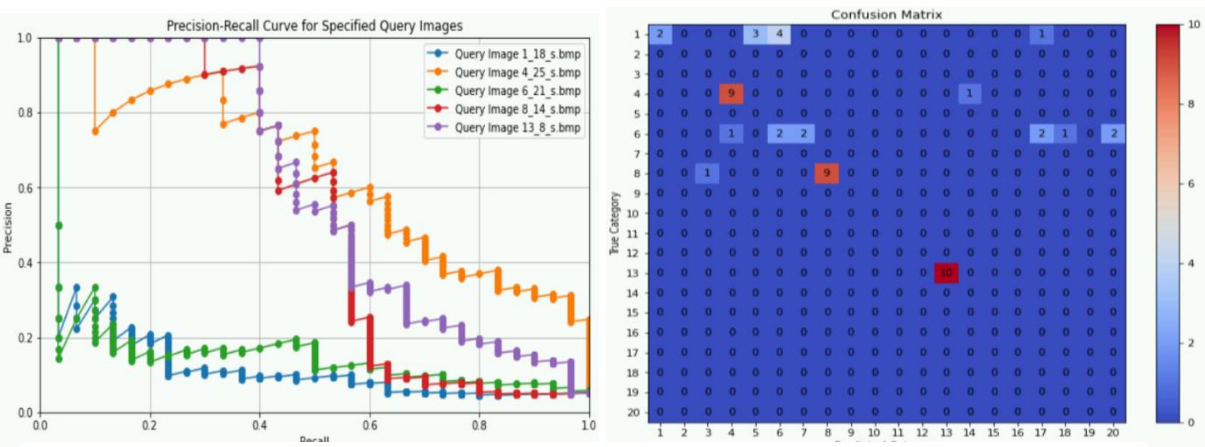

Figure 9 : Precision-Recall curve and confusion matrix for the SIFT descriptor, showing the retrieval performance across 5 query images.

Results

The SIFT detector outperforms other detectors, especially in categories like 4, 8, and 13 (Figure 9), where it shows marked improvements. The precision-recall (PR) curve for these categories indicates better retrieval accuracy, with the lines positioned higher, suggesting that more relevant images are retrieved. This performance is further confirmed by the confusion matrix, where the SIFT model registers more true positives in these categories. For specific objects like books, aeroplanes, and bicycles, the detector demonstrates high accuracy in retrieving correct images, as seen in the visual outputs.

However, for categories 1 (cow and grass) and 6 (human), performance remains relatively low (Figures 8 and 9). This is partly due to a lack of contextual understanding; the SIFT detector focuses primarily on visual features like color and texture, which can lead to confusion when objects share similar attributes or background elements. There is also object overlap issue category 6 and 19 are same but differently defined in the dataset. Improving dataset may help this detector to perform well on these categories.

The SIFT detector's strong performance can be attributed to its superior feature extraction through the Bag of Visual Words approach and its effective clustering of visual words with K-means. This enables the model to distinguish objects more accurately within challenging categories, creating distinct clusters for better classification and retrieval. These improvements contribute to its robust handling of various categories and the ability to retrieve images with greater precision and clarity.

Conclusions

The evaluation of descriptors, distance metrics, and detectors highlights key strengths and challenges related to the dataset and task. The global color histogram works well for distinct color regions but struggles with categories where objects share similar colors or appear against complex backgrounds, such as cow and grass or human. The spatial grid histogram, enhanced by PCA, improves performance by incorporating texture and reducing noise, especially helpful for categories with intricate textures. Using Mahalanobis distance with the spatial grid further enhances feature separation, boosting performance.

The SIFT detector achieves high precision-recall values for visually distinct categories, such as books, aeroplanes, and bicycles, yet faces misclassification challenges in overlapping categories like human" and animal, reflecting dataset limitations. Euclidean distance performs reliably across descriptors, especially with the spatial grid, while Mahalanobis distance with PCA provides efficient feature separation. Other metrics, like Manhattan distance and cosine similarity, show lower accuracy in retrieval and classification.

A notable challenge for the descriptors is focusing on the correct object in multi-object images, impacting precision-recall performance. Overall, combining spatial grid histograms with PCA and Mahalanobis distance offers a robust solution for datasets with color and texture complexities, while the SIFT detector remains effective for visually distinct objects.

References

1. Dataset msrc_objcategimagedatabase_v2.zip
2. R. Szeliski, *Computer Vision: Algorithms and Applications*, 1st ed. New York, NY, USA: Springer, 2010.
3. J. Narwade and B. Kumar, "Local and Global Color Histogram Feature for Color Content-Based Image Retrieval System," in Advances in Computer Science and Information Technology, 2016, pp. 293-300. doi: 10.1007/978-981-10-0767-5_32.

4. J.-C. Chen, C.-C. Chen, and C.-H. Chuang, "Intelligent Image Retrieval Using Texture and Color Features," in 2019 IEEE International Conference on Consumer Electronics - Taiwan (ICCE-TW), Yilan, Taiwan, 2019, pp. 1-2. doi: 10.1109/ICCE-TW46550.2019.8991959.

5. O. A. Adegbola, I. A. Adeyemo, F. A. Semire, S. I. Popoola, and A. A. Atayero, "A Principal Component Analysis-Based Feature Dimensionality Reduction Scheme for Content-Based Image Retrieval System," Journal of Computer Science, vol. 11, no. 6, pp. 1-7, 2015.

6. A. A. Khodaskar and S. A. Ladhake, "Pattern Recognition: Advanced Development, Techniques and Application for Image Retrieval," in 2014 International Conference on Communication and Network Technologies, Sivakasi, India, 2014, pp. 74-78. doi: 10.1109/CNT.2014.7062728.