ete technology

Building an Image-Based Place Recognition & Retrieval System

Candidate Name: Betül USLU

betuluslu5u5@gmail.com

1. Overview

This project implements an end-to-end image-based place recognition and retrieval system that identifies the most visually similar location for a given query image using deep visual embeddings and cosine similarity search. The system is designed to be modular, reproducible, and computationally efficient, while also supporting open-set recognition for unknown locations that do not exist in the gallery.

The full pipeline consists of dataset validation, embedding extraction, index construction, similarity-based retrieval, open-set handling, and quantitative evaluation using standard retrieval metrics such as Recall@K and Mean Average Precision (mAP).

2. Dataset Validation and Integrity Checks

Before building the retrieval system, the dataset was validated to ensure structural correctness and data reliability. The dataset is organized using a `manifest.csv` file with two splits: gallery (reference images) and query (search images).

A dedicated validation script was executed to detect:

- Missing image files
- Corrupted or unreadable images
- Extremely small resolution images
- Cross-split duplicates (data leakage risk)

Validation Results:

- Total Records: 66
- Gallery Images: 37
- Query Images: 29
- Missing Files: 0
- Unreadable Images: 0
- Too Small Images: 0
- Cross-Split Duplicates: 0

These results confirm that the dataset is clean and suitable for a fair retrieval evaluation without leakage or corrupted samples.

3. Embedding Extraction Strategy

For visual feature extraction, a pretrained ResNet50 model was used as a fixed backbone. The classification head was removed and the model was used purely as a feature extractor, producing 2048-dimensional embeddings for each image.

ResNet50 was selected because it provides a strong and reliable baseline for visual representation learning while remaining computationally efficient. Since it is pretrained on ImageNet, it captures high-level semantic features such as structure, texture, and scene composition without requiring additional training. This makes it especially suitable for a lightweight retrieval pipeline where training is not the primary focus.

Another important advantage of ResNet50 is its balance between accuracy and inference speed. Compared to heavier models such as Vision Transformers or CLIP, ResNet50 can run efficiently on CPU while still producing stable and discriminative embeddings. This aligns well with the goal of building a reproducible and resource-efficient system.

During embedding extraction:

- Images were converted to RGB when necessary
- Preprocessing followed standard ImageNet normalization
- Batched inference was used for efficiency
- L2 normalization was applied to all embeddings

L2 normalization is critical because it ensures that cosine similarity can be computed efficiently using a dot product while keeping similarity scores stable across different images.

To improve reproducibility and runtime efficiency, extracted embeddings were cached as `.npy` files. The logs confirm that cached embeddings were successfully reused, preventing redundant model inference and significantly speeding up the pipeline.

4. Index Construction and Representation

After extracting gallery embeddings, a compressed index file (`index_resnet50.npz`) was created containing:

- Gallery embedding matrix (N = 37, D = 2048)
- Corresponding image paths

This indexing step allows the system to perform fast similarity search without recomputing embeddings, which is a common design choice in real-world retrieval systems.

5. Similarity Search Methodology

The retrieval stage is based on cosine similarity between query and gallery embeddings. Since all embeddings are L2-normalized, cosine similarity becomes equivalent to a dot product, enabling fully vectorized and efficient computation using NumPy.

For each query image:

1. The similarity score with all gallery embeddings is computed
2. Scores are sorted in descending order

3. Top-K most similar images are returned

Example retrieval output shows that the top-ranked results belong to the same landmark class, indicating that the embedding space successfully captures semantic similarity rather than low-level pixel similarity.

This approach is simple, scalable, and widely used in modern image retrieval systems due to its computational efficiency and robustness.

6. Open-Set (UNKNOWN) Recognition Approach

A key requirement of the system is handling open-set queries, where some query images may not correspond to any location in the gallery. To address this, a threshold-based UNKNOWN detection mechanism was implemented.

The decision rule is defined as:
If Top-1 cosine similarity < τ → Predict UNKNOWN

The threshold value τ was not chosen arbitrarily. Instead, it was determined empirically by analyzing the similarity score distribution during evaluation. It was observed that:

- Known queries consistently produced higher similarity scores
- Open-set queries produced noticeably lower Top-1 similarity scores

Different threshold values (0.60, 0.55, and 0.50) were tested. Higher thresholds (e.g., 0.60) led to over-rejection, incorrectly labeling valid matches as UNKNOWN. Lower thresholds risked accepting incorrect matches. The value $\tau = 0.50$ provided the best balance, achieving perfect separation between known and unknown queries on the dataset.

Final open-set performance:

- Open-set Queries: 9
- Predicted UNKNOWN: 9
- False Positives: 0
- False Negatives: 0
- UNKNOWN Precision: 1.00
- UNKNOWN Recall: 1.00

This demonstrates that the selected threshold is both stable and data-driven.

7. Evaluation Methodology and Metric Design

The system was evaluated using Recall@K (K = 1, 5, 10) and Mean Average Precision (mAP). Unlike single-label classification, place recognition is a multi-positive retrieval problem because each location may have multiple correct gallery images.

Therefore, the evaluation was designed at the item level:

- All gallery images from the same location are treated as positives
- Ranked retrieval lists are compared against multi-positive ground truth
- Open-set queries are excluded from Recall and mAP to ensure fair evaluation

The evaluation metrics were also validated using unit tests to ensure mathematical correctness and reliability.

8. Quantitative Results

The final evaluation was performed using an UNKNOWN threshold of 0.50.

Results:

- Total Queries: 29
- Open-set Queries: 9
- Recall@1: 1.00
- Recall@5: 1.00
- Recall@10: 1.00
- mAP: 0.9019
- UNKNOWN Precision: 1.00
- UNKNOWN Recall: 1.00

A Recall@1 score of 1.00 indicates that the correct location was retrieved as the top result for all valid queries. The mAP score of approximately 0.90 shows strong ranking quality across multiple positive matches per location. The perfect UNKNOWN precision and recall further confirm the robustness of the open-set detection mechanism.

9. System Reliability and Testing

To ensure correctness, unit tests were implemented for metric functions, including Recall@K and Average Precision calculations. All tests passed successfully, confirming that the evaluation logic is consistent and mathematically sound.

Additionally, caching mechanisms, modular scripts, and structured logging improve reproducibility and maintainability of the pipeline, which are essential qualities in production-grade retrieval systems.

10. Conclusion

An end-to-end image-based place recognition and retrieval system was successfully designed and implemented using pretrained deep visual embeddings and cosine similarity search. The system demonstrates strong retrieval accuracy, robust open-set handling, and reproducible performance through a modular and well-validated pipeline.

The use of ResNet50 provided stable and semantically meaningful embeddings without requiring additional training, while the empirically selected threshold ($\tau = 0.50$) enabled accurate UNKNOWN detection without over-rejecting valid matches. Overall, the results show that the system is reliable, efficient, and well-aligned with real-world visual retrieval requirements.