Vision and Image Processing:
Assignment 4: Content Based Image Retrieval

Mostafa Elshamy, Tom Vig, Andreas Hammer

January 2025

1 Introduction

In this report, we attempt to implement a prototypical Content-based image retrieval (CBIR) system. This mean that the system doesn't rely on metadata for the image retrieval - but rather on features extracted from the content of the image (using SIFT).
We evaluate the results of our system using 2 different evaluation methods (MRR, top-3 accuracy) in two separate experiments outlined below.

2 Methodology

We conducted two experiments to observe the accuracy of the implemented Content Based Image Retrieval. For the first, we retrieve images from the training category, and for the second, we retrieve images from the test categories. For each experiment, we ran tests on a sample of five randomly selected categories, then expanded to 20 categories. This method was chosen because it allows for a more clear graphical representation of the results when displaying 5 categories, and more accurate results when computing the aggregates in 20 categories. For both experiments, similar images are retrieved using the Common Words method of CBIR. To determine the number of clusters to be used in the classification model, we ran tests for a range of k clusters, and measured the compactness of the results by measuring the average distance from the center of each cluster to the points assigned to that cluster. The results are shown in figure 1, using the 'elbow method', we determined that using a k value of 300 was the most optimal compromise between result accuracy and computational efficiency.

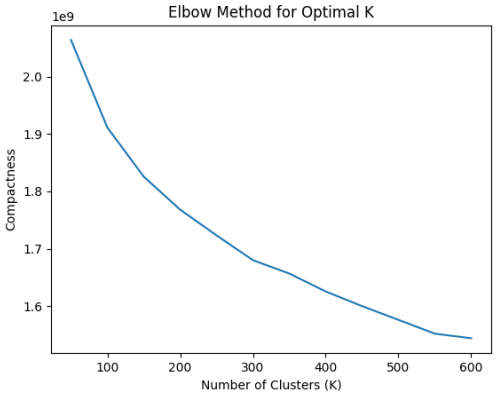


Figure 1: Elbow Graph for optimal clustering

3 Experimental Results

3.1 Retrieving Training Images

In this experiment we attempted to retrieve training images. Because each query image is taken from the training data and the retrieved results are also from the training data, it means that the query image itself will always be the top result, with 1.0 similarity. This means that in both the MRR and Top 3 accuracy metrics we will get perfect results. In order to achieve more meaningful results, we decided to exclude the first image result from the retrieved images, and to start counting from the second result.

Sampled Categories Figures 2 and 3 show the results of this experiment on a sample of 5 categories. Figure 2 shows the Mean Reciprocal Rank of the results. This metric is calculated by finding the first image of the returned images that matches the category of the queried image, and reciprocates the order in which it appears in the list of most similar images.
Figure 3 presents the ratio of instances that a queried image results in a correct categorization within the top 3 most similar images, as a percentage.

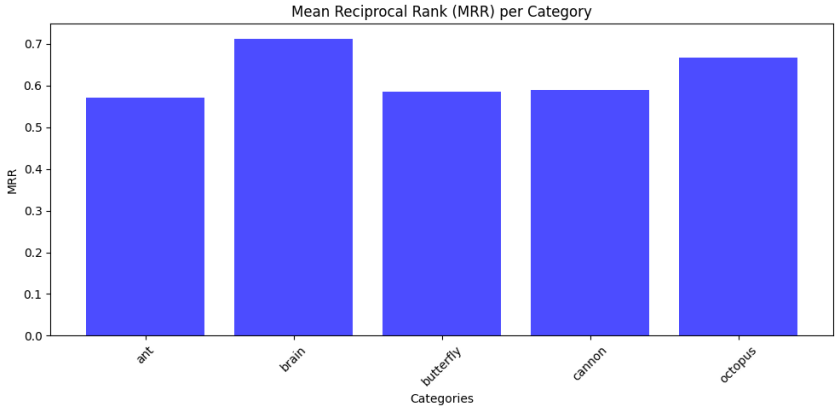


Figure 2: MRR 5 Categories

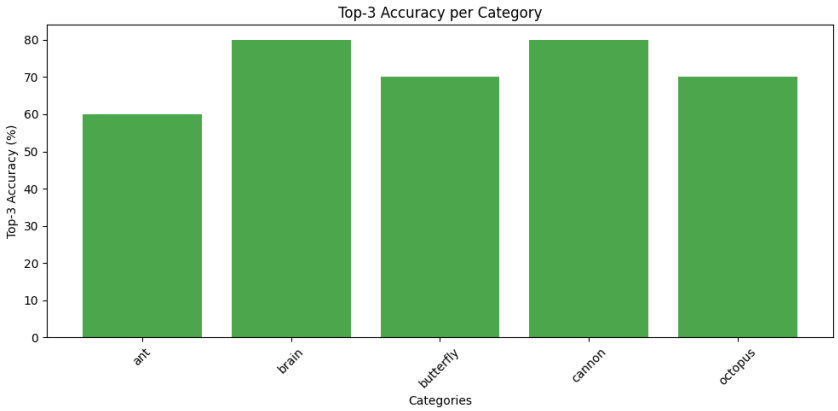


Figure 3: Top 3 Images for 5 Categories

All Categories After testing the CBIR algorithm on a sample of half the images in only 5 categories, we expanded the experiment to include all images across all 20 categories. The results of this are shown in figure 1. The overall accuracy is 39.47% and average MRR is 0.3597. For both metrics, the standard deviation between categories is low.

Category	MRR	Top-3 Accuracy
ant	0.3889	50.00%
brain	0.5373	60.00%
buddha	0.3206	40.00%
butterfly	0.3472	30.00%
camera	0.6098	70.00%
cannon	0.4388	50.00%
cellphone	0.1705	20.00%
chair	0.2053	10.00%
crocodile	0.2321	20.00%
dolphin	0.4390	50.00%
elephant	0.2548	20.00%
emu	0.8643	90.00%
flamingo	0.1782	20.00%
headphone	0.2286	40.00%
lamp	0.1416	20.00%
lotus	0.2399	40.00%
menorah	0.3041	30.00%
octopus	0.6111	60.00%
pigeon	0.3229	30.00%
Overall	**0.3597**	**39.47%**
σ	**0.1871**	**0.2068**

Table 1: Results of Retrieving Training Images from 20 Categories

(a) Query Image **(b)** 0.611374 **(c)** 0.595738 **(d)** 0.587889

Figure 4: Accurate Result

An example of a relatively accurate query result is shown in Figure 4.The queried image (a) belongs to the emu category of images, and the top two out of 3 of the returned images also belong to that category. The third image returned is a cannon, but by observing the image it is clear why the common words algorithm chose this to be similar to the queried image, as they both contain a round body, and a long neck pointing downwards. The similarity index of each image is captioned beneath each one.

For contrast, consider figure 5. The queried image (a) is a chair, but none of the top 3 most similar images are chairs. The inaccuracy may have occurred due to all images having a similar light colored background, and a single object in the middle.

3.2 Retrieving Test Images

This experiment consisted of 2 parts, similar to the first, except now we used half the data to train the model, and the other half as testing images. An image from the test set is provided as a query, and the top 3 most similar images from the test set are returned. This experiment provides a more accurate use-case scenario for image search engines like google-image search, as both the queried image and the set of images to test have not been used to train the model.

(a) Query Image **(b)** 0.527943 **(c)** 0.518095 **(d)** 0.491115

Figure 5: Inaccurate Result

Sampled Categories Starting initially with the same 5 categories used in the previous experiment, figures 6 and 7 show the MRR and Top-3 accuracy respectively. The results of these categories are in general lower than those of the previous experiment, with the most significant result being that none of the top 3 images returned by querying octopus images returned an image of that category. The MRR for the octopus category is 0.1, meaning that it was not until the tenth most similar image that a correct categorization appeared.

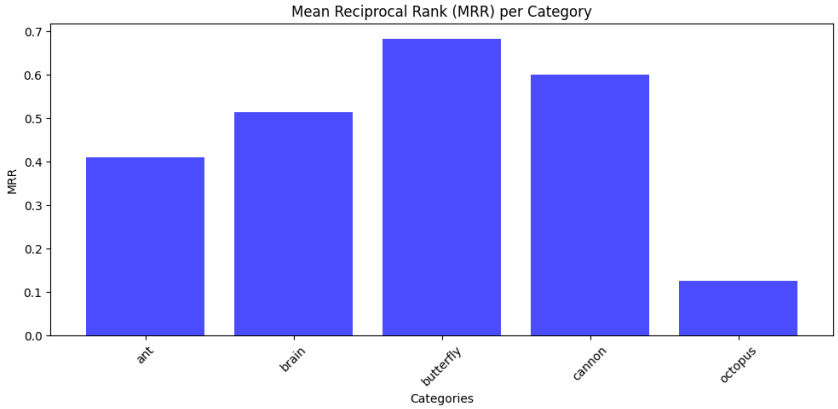


Figure 6: MRR 5 Categories

All Categories Expanding our experiment to use all images in the 20 categories returned mixed results, as can be seen in table 2. For certain categories, high levels of accuracy were achieved, such as emu and headphones achieving 90% and 70% respectively. However, the octopus category continued to struggle, with its MRR dropping to 0.041. Overall, our implementation of Content Based Image Retrieval resulted in 35.79% accuracy in returning majority correct images, and on average the third most similar image is the first to be of the correct category. The average classification accuracy for the test images all categories was 20.53% (where classification was decided by the most common category appearing in the top 5 results).

Taking a look at specific examples gives some insight into why this might be. Figure 9 shows the result of querying an image of the octopus category. None of the given results are octopi, however the two most similar images (b and c) are ants. The ants, similar to the octopus, are made up of a round body with long thin lines protruding from it. This could explain why these categories became so intertwined (and hence inaccurate). By contrast, figure 8 shows a more accurate result for the cannon query.

4 Conclusions

In this report, we explored the development and evaluation of a CBIR system using SIFT feature extraction. The experiments demonstrate the system ability to retrive images based on content

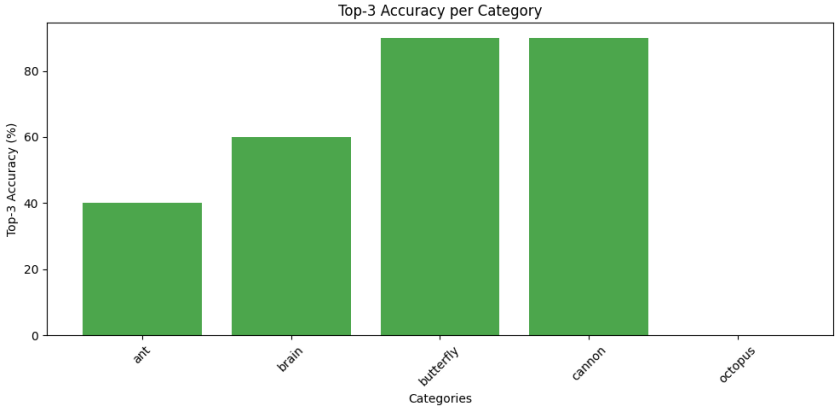


Figure 7: Top 3 Images for 5 Categories

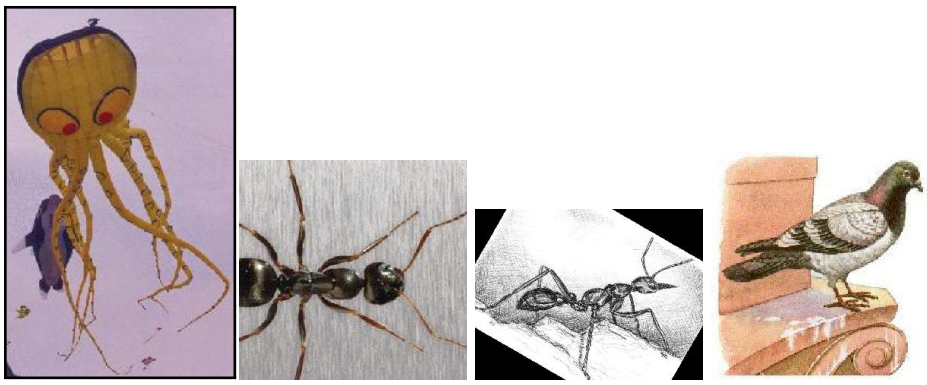
(a) Query Image **(b)** 0.631137 **(c)** 0.576323 **(d)** 0.551990

Figure 8: Accurate Result

similarities, providing some understanding of its strengths and weaknesses.

The results from the first experiment, which uses training images as queries, showed promising results in both the MRR and top-3 accuracy metrics. However, moving to the testing images highlighted the challenges of generalizing to unseen data, as there was a drop in the quality of the retrieved results. This shows the importance of the choices in the feature extraction and clustering process, and suggests that there might be a need for advanced techniques such as deep learning-based approach.

Through the experiments conducted, we observed that some categories are easier to identify than others. For example, "emu" and "camera" both consistently yielded relatively high retrieval accuracies, likely due to their distinct features. On the other hand, categories such as "chair" proved to be a challenge, as the system struggled to differentiate images with similar backgrounds or textures. The significant decline in the "octopus" category between the training and testing results showcased the importance of training on a large and varied image dataset.



(a) Query Image **(b)** 0.519375 **(c)** 0.518729 **(d)** 0.516954

Figure 9: Inaccurate Result

Category	MRR	Top-3 Accuracy
ant	0.3692	40.00%
brain	0.2770	20.00%
buddha	0.2294	20.00%
butterfly	0.4661	60.00%
camera	0.4728	60.00%
cannon	0.6083	60.00%
cellphone	0.1805	20.00%
chair	0.1590	20.00%
crocodile	0.3144	40.00%
dolphin	0.3498	30.00%
elephant	0.3539	40.00%
emu	0.7458	90.00%
flamingo	0.0400	00.00%
headphone	0.5511	70.00%
lamp	0.1193	10.00%
lotus	0.4338	60.00%
menorah	0.2130	30.00%
octopus	0.0410	00.00%
pigeon	0.1246	10.00%
Overall	**0.3184**	**35.79%**
σ	**0.1946**	**0.2524**

Table 2: Results of Retrieving Training Images from 20 Categories

When comparing the results of the experiments with 5 categories or 20 categories, we observed very similar trends - with slight variations. The overall accuracy and ranking quality remained relatively stable between the two choices of categories, which suggests that the system is not overly sensitive to the number of categories but rather to the distinctiveness of features within each category.