Assignment 4

Amalia Voulgaraki - nbt146, Vasiliki Kouselini - tbm262,
Arina Zamyshevskaya - tjw488, Daria Radcenko - lzw942

5th January 2026

1 Data Set

For this assignment, we used a subset of the Caltech101 data set. This image set consists of 101 categories, where each one is made up of 31 to 800 images [1]. Five sample images can be seen in Figure 1.

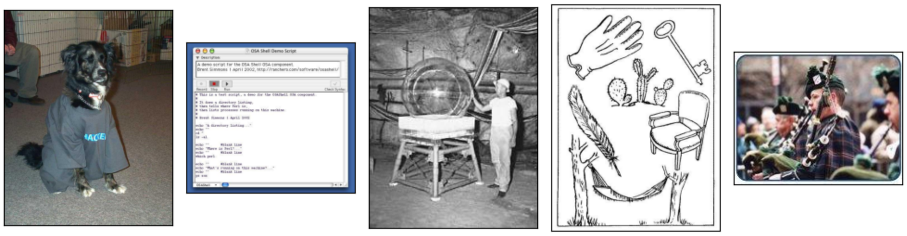

Figure 1: Five sample images of the Caltech101 data set.

To reduce computational power, we extracted the first 20 categories of the data set and used all images available. Then, we shuffled the images to create a random order and split the data into a training and testing set of equal size.

2 Codebook Generation

During codebook generation, we first used scale-invariant feature transform (SIFT), which is an algorithm that extracts the important features of an image (e.g. edges, corners, and texture regions). We detect important keypoints and later match them across images in classification. First, we converted the images into grayscale and performed SIFT afterwards, as most SIFT version do not include color, which could impact the results negatively. We obtain a matrix, where one row consists of one descriptor.

Figure 2 shows five sample images with their corresponding keypoints. Each circle is one keypoint, where the size of the circle represents the scale at which it was detected. The lines in the circles indicate the dominant direction of the gradient. Different features are represented by different colors.

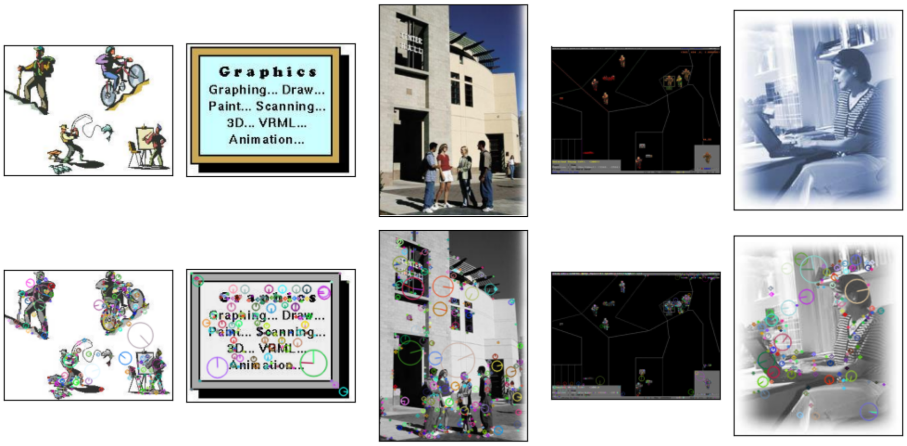

Figure 2: Five sample images of the Caltech101 data set with their corresponding keypoint maps.

To obtain the optimal codebook size, we performed cross-validation on the training data and investigated both mean reciprocal rank (MRR) and top-3 accuracy. For this, we looked at values in the range of 700 to 2000 in steps of 300. As validating on one tenth of the data would not have been significant enough, we reduced the number of folds to five.

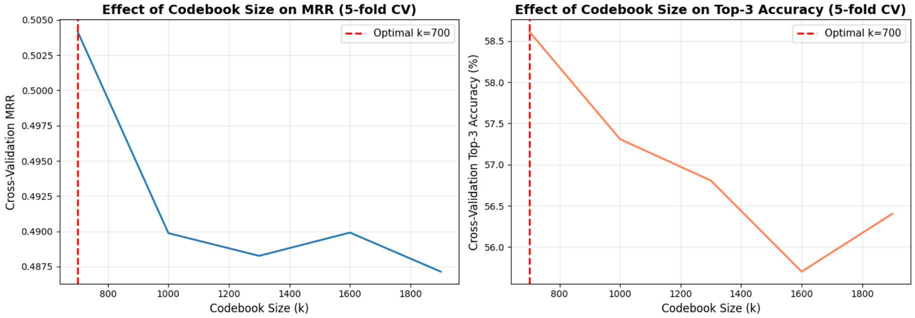

Figure 3: The effect of k on MRR (left) and top-3 accuracy (right) during five-fold cross validation.

As shown in Figure 3, a lower amount of clusters overall tends to produce minimally better results than a high amount of clusters, although the effect does not seem to be significant. In our case, a k of 700 yields the best results for both MRR and top-3 accuracy.

After obtaining the optimal parameters, we perform k-means on the descriptors we extracted during SIFT with k=700. The clustering algorithm groups these descriptors in such a way that similar features lie in the same cluster, while the distance to very dissimilar features is maximized. This step subsequently helps us to create a bag of visual words that can be used for image classification later.

3 Results

To evaluate our bag of visual words representation, we take both average MRR and top-3 classification accuracy for the training data and the testing data respectively. Furthermore, we looked at both metrics for each class for more detailed insights.

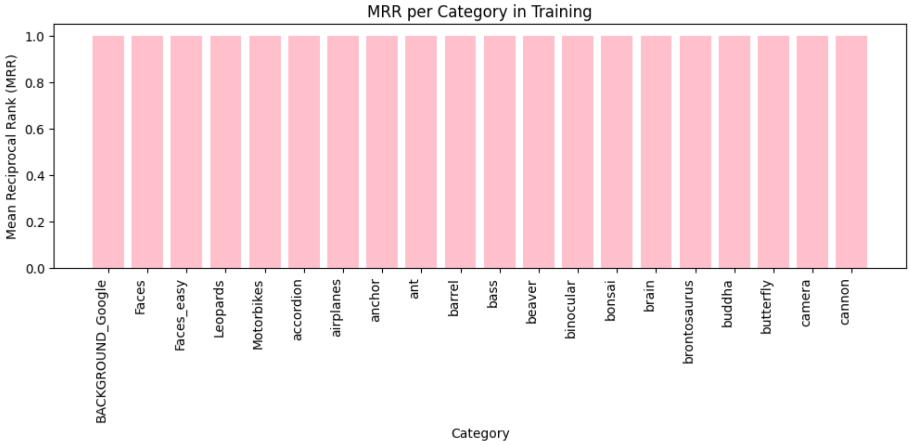

Figure 4: MRR for all 20 categories in the training data set.

As seen in Figure 4, each category has an MRR of 1, leading to a perfect score when retrieving training data.

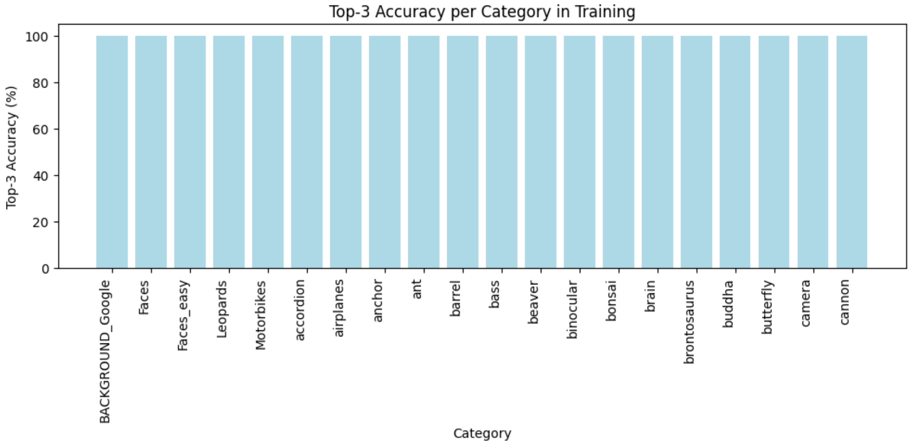

Figure 5: Top-3 accuracy for all 20 categories in the training data set.

We observe the same results for the top-3 accuracy in Figure 5, where individual and overall scores are 100%.

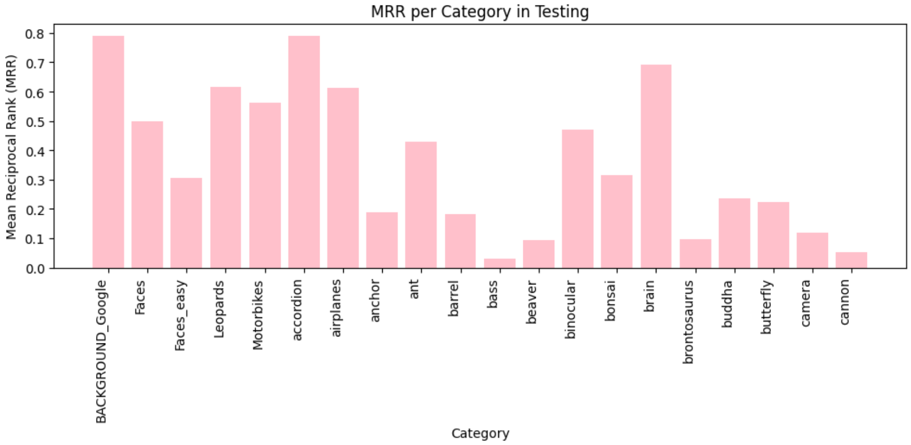

Figure 6: MRR for all 20 categories in the test data set.

In contrast to the results above, the overall MRR in the testing set is 0.5112 with a lot more variety between individual categories. As seen in Figure 6, BACKGROUND_Google and accordion yield the highest MRR at around 0.8, whereas the MRR for brass lies at approximately 0.05.

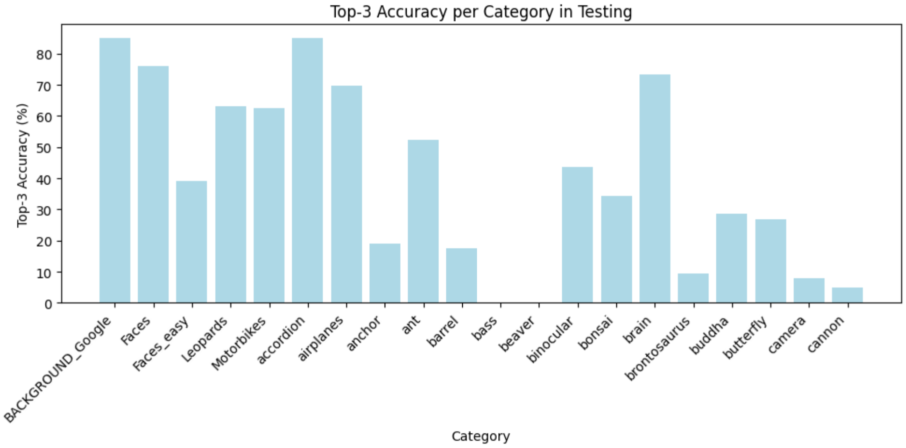

Figure 7: Top-3 accuracy for all 20 categories in the test data set.

Looking at the top-3 accuracies for the testing data set in Figure 7, we observe the same overall pattern as in Figure 6, but with slightly more pronounced differences between the categories. This leads to a moderately higher top-3 accuracy of 59.02%. As before, BACKGROUND_Google and accordion show the highest top-3 accuracy at around 80%. Both bass and beaver have not had a single correctly classified image in the top three.

4 Reflection

Analyzing our results, the large gap between training and testing accuracy suggest that adjustments are needed to get a more optimal outcome. We identified multiple possible reasons for this generalization gap.

One contributing factor could be the high variability of the number of images in each category. A few categories, such as airplane (800) and BACKGROUND_Google (468) have a significantly higher amount of images compared to categories like bass (54) and beaver (46) [1]. A possible consequence could be that the larger represented categories are classified correctly more frequently than the categories with less images. Furthermore we noticed that in some categories, such as Faces, similar images tend to repeat themselves, while other classes, like buddha, show more variability. In the future, we might consider a more balanced data set.

Another possible reason could be our chosen cluster size (700). As seen in Figure 3, the optimal k in the range of 700 to 2000 seems to be 700. We have not cross-validated for values smaller then 700. However, the trajectory of the line suggests, that the real optimal k could be smaller than our chosen value.

This could lead to overfitting on the training data and ultimately explain the observed gap in training and testing accuracy. The amount of clusters could be too high for the amount of descriptors we extract, therefore leading to the memorization of images rather than the learning of features. An optimal k could prevent this by leading to a more generalizable fit.

References

[1] Bikram Saha. Caltech-101 dataset. `https://www.kaggle.com/datasets/imbikramsaha/caltech-101`, n.d. Retrieved January 4, 2026.