Assignment 1

5^{th} February 2019

1 Problem Statement

Given a dataset of images of several objects our job was to train a model on it such that it is able to retrieve all the instances of a given test object.

2 Methods Used

2.1 Method 1

We approached the problem using the bag of visual words model.The general idea of bag of visual words is to represent an image as a set of features. Features consists of keypoints and descriptors. Keypoints are the "stand out" points present in the image and descriptors are the description of keypoints. We made use of keypoints and descriptors to represent the image as a frequency histogram of the features that are present in the image.

We used the SIFT algorithm for feature extraction to build a visual dictionary. Next we clustered these features using KMeans algorithm. The centres of each cluster are our visual words. Then we built a frequency histogram for each image in the training set using the visual words.

We also built the tf-idf vector for each image.We represent each image using k terms $t_1, t_2, ..., t_k$ where t_i is given as:

$$t_i = \frac{n_{id}}{n_d} \log(\frac{N}{n_i}) \tag{1}$$

Here n_{id} is the number of occurrences of word i in document d , n_d is the number of words in document d , n_i is the number of term i in the whole database and N is the number of documents in the database.

We calculated a similar histogram for the each test image and then calculated the similarity of test image with each training image using cosine similarity. Here the tf-idf values were used as a weight while calculating the cosine similarity.

Using a stop list analogy we removed the most frequent visual words that occur in almost all images.

This method worked fairly well when there was no background noise but in presence of background noise it performed poorly.

2.2 Method 2

We also used a deep learning model for the task. We used a Convolutional Neural Network for classification. The data consists of images belonging to 16 different classes. Our goal was to get a ranking of the 16 classes representing the likelihood of the test image belonging to a particular class. We assigned labels for each image in the training data. The model outputs a 16 length score vector denoting the likelihood of each class. The model was trained on the training data and the loss function used was softmax. If label for i-th image is j, and y_k denotes the score of the k-th class, then the loss is:
$L = -\log \left[\frac{e^{y_j}}{\sum_{k=1}^{16} e^{y_k}} \right].$
For each test image we considered the classes in the decreasing order of scores and returned the images from that class. The preformance of this method was poor since there was a lot of difference between the training and test images.

2.3 Method 3

We also used DELF (DEep Local Features) which is a DL model that achieves state of the art image retrieval results.It works by extracting key points in an image and descriptions for those key points. At retrieval time, it'll find images with similar descriptions, and matches their key points geometrically to verify if the image is a proper match. The important difference between this approach and the one using SIFT is that it also performs geometric verification using RANSAC which make sure that matches are consistent with a global geometric transformation.This method gave much better results as compared to the bag of visual words model. It gave correct results even in the presence of background noise.

3 Challenges Faced

We faced a number of challenges during the implementation of our algorithms.One problem was that all the training images had a chessboard pattern in them and most of the keypoints of the SIFT detector would lie on the chessboard pattern. We tried to solve this problem by cropping the training images before extracting features. This was possible since almost all the training images had the objects in the centre. We also suppressed the most frequently occurring visual words to solve this problem. Another problem which we faced was that when the background of the test image was not white then the SIFT detector was not able to detect correct keypoints.Thus our algorithm failed to give correct answer for these kind of test images. The DELF algorithm was able to overcome this problem and gave correct matches. The problem that we faced in solving the problem using deep learning method was that the training and test images were quite different from each other.So although we could get high accuracy on the validation set we could not get a good result on test set.

4 Saved Models

The MD5 hashes of the models that we saved are:

histogram.obj	3f36019ed3a6a6896dddd2bed610eda3
stop_words.obj	4a5c5ea68db280c825a95e80062eb88f
cluster_centers.obj	7072ecd9a6562cf859c905cdbfb40db1
tf_idf.obj	95d9aca7dbf80eaac1e47100b24d213e
dict.obj	0053c81ff51f20b0dba107bfabe23327
results.obj	3020fae4acf24d4225492324a2e962a0