Capstone Proposal: Social Image Description Platform

Xing Hao

Oct. 15th, 2017

Domain Background

Computer vision tasks include methods for extracting features, recognizing, classifying images, and more. Since the appearance of Artificial Intelligence, scientists have spent a lot of effort on this topic.

There are several techniques to solve the computer vision problems, including:

- SIFT and SURF for feature-point extraction which can be used for object recognition.
- Viola-Jones algorithm, for object (especially face) detection in real time.
- 'Eigenfaces' approach, using PCA for dimension reduction. This algorithm is used in face recognition.
- Machine learning algorithms, such as Neural Networks for image classification.
- Lucas-Kanade algorithm, Mean-shift algorithm, and Kalman filter for object tracking.

This project will focus on object recognition and image description which has a lot of applications such as image searching, products recommendation, and so on.

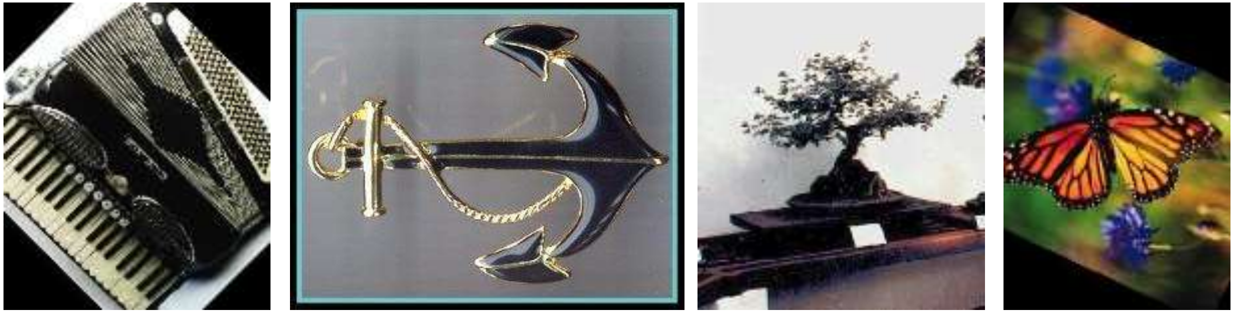
This problem can be solved using SIFT (Scale-invariant feature transform) algorithm [1]. SIFT algorithm is trying to extract local features of objects from a set of reference images and store the local features in a database. An object is recognized in a new image by individually comparing each feature from the new image to this database and finding candidate matching features based on distance of their feature vectors. Feature vectors can also be used to do image comparison and classification.

Problem Statement

This project is trying to build a social image description platform. Users can upload images. The platform will extract features of images, use them to classify images, and predict the descriptions for the images.

The input of the machine learning task is an image. The output is the description of the image which will includes the top 5 labels which match the images with higher *confidence* which will be explained later. The platform will also show the most similar images.

Datasets and Inputs



accordion anchor bonsai butterfly

Figure 1. Example images

Regarding the training set, I will use the images of Caltech 101[2] which includes objects belonging to 101 categories. About 40 to 800 images per category. Most categories have about 50 images. The size of each image is roughly 300 x 200 pixels. The categories are accordion, anchor, bonsai, brain, butterfly, and more. Figure 1 shows some example images in the data set.

Solution Statement

I will use SIFT the extract local features for each image, and for all the local features of images in the same category, use k-means to cluster local features. Use the centers of all the clusters as the vocabularies of this category, and use them to compute the BOW of each image. Using the BOW as the input, train SVM to classify images. Each category has an SVM where the label of the images in this category is 1, and all the other images are 0.

Benchmark Model

The dataset includes 8677 images, and the largest category has 800 images, so the accuracy of a model that always predict the majority class is 9.2%. I will calculate the accuracy for each category and compare it with 9.2%.

Evaluation Metrics

The Caltech 101 dataset will also be used as the test set. If the true category of the image is in the description, then the classification is marked as correct. The accuracy of the platform is calculated as (correct images)/(all images).

Project Design

1. Training

The training phase includes 4 steps including features extraction, clustering, and SVM training, as shown in Figure 2:

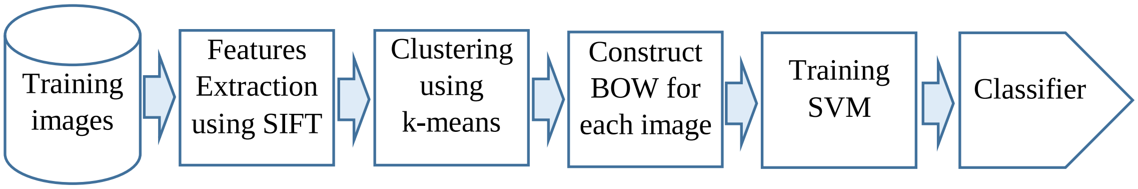


Figure 2. Training

- **Features Extraction using SIFT**
  SIFT algorithm is trying to extract local features. Using SIFT, the local features of each image will be extracted. Each feature will be a N*M matrix where N is the number of local features, and M is the dimension of one feature.
- **Clustering using k-means**
  Using the local features from step 1 as the input, clustering the local features for each category. Using the centers as the vocabularies of the category.
- **Construct BOW for each image**
  For each local feature of an image, find the nearest center which will be the word for this image. Count the number of each center and this is the bag of words of this image.

- **Training SVM**
  Using the BOW of all the images, train SVM. For each category I will train a binary SVM where the label of the images in this category is 1, and all the other images are 0. There will be one SVM for each category.

2. Testing

For each image, using all the SVM to calculate the label and the distance to the margin. Using the label (1 or 0) and the distance to sort the images, and return the top 5 categories as the description of the image. For each category, the *confidence* of the category is calculated as (*label*-0.5)*2**distance*. Return the 5 categories with the highest confidence as the description of the image. In the 5 categories, find the nearest neighbors of the image, and return them as the similar images.

Reference

[1] Lowe, David G. "Object recognition from local scale-invariant features." In *Computer vision, 1999. The proceedings of the seventh IEEE international conference on*, vol. 2, pp. 1150-1157. Ieee, 1999.

[2] Caltech 101: http://www.vision.caltech.edu/Image_Datasets/Caltech101/