Statistical Methods in AI

Project Report

Project Title: Object Recognition

Team: HGP

Smriti Singh (201301080), Harshit Harchani (201301105), Vivek Ghaisas (201301106)

Dataset Description

Link: http://www.vision.caltech.edu/Image_Datasets/Caltech101/
Caltech 101 is a data set of digital images created in September 2003 and compiled by Fei-Fei Li, Marco Andreetto, Marc 'Aurelio Ranzato and Pietro Perona at the California Institute of Technology. It is intended to facilitate Computer Vision research and techniques and is most applicable to techniques involving image recognition classification and categorization. Caltech 101 contains a total of 9,146 images, split between 101 distinct object categories (faces, watches, ants, pianos, etc.) and a background category. Provided with the images are a set of annotations describing the outlines of each image, along with a Matlab script for viewing.

- The images are cropped and re-sized.
- Many categories are represented, which suits both single and multiple class recognition algorithms.
- Detailed object outlines are marked.
- Available for general use, Caltech 101 acts as a common standard by which to compare different algorithms without bias due to different data sets.

Techniques

This is a script for content based image classification using the bag of visual words approach.
→ **SIFT (Scale Invariant Feature Transform)** to detect and describe local features of an image. For any object in an image, interesting points on the object can be extracted to provide a "feature description" of the object. This description, extracted from a training image, can then be used to identify the object when attempting to locate the object in a test image containing many other objects. To perform reliable recognition, it is important that the features extracted from the training image be detectable even under changes in image scale, noise and

illumination. Such points usually lie on high-contrast regions of the image, such as object edges. SIFT can robustly identify objects even among clutter and under partial occlusion, because the SIFT feature descriptor is invariant to uniform scaling, orientation, and partially invariant to affine distortion and illumination changes. This returns a set of features per image; the number of features per image is variable.

→ **Mini Batch K-means** for clustering of the features generated for the images.
Its main idea is to use small random batches of examples of a fixed size so they can be stored in memory. Each iteration a new random sample from the dataset is obtained and used to update the clusters and this is repeated until convergence. In our script, we run the algorithm to compute 512 clusters out of all the features of all the images given to the script for training. Based on these these 512 clusters, an image can be classified into a codeword, that maps an image to the clusters.

→ **SVM** for classification of the images, based on these features.
Support Vector Machines are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis. Given a set of training examples, each marked for belonging to a categories, an SVM training algorithm builds a model that assigns new examples into any one category. SVMs can efficiently perform a non-linear classification using what is called the kernel trick, implicitly mapping their inputs into high-dimensional feature spaces. The SVM model trains on the classes and the codewords of the images corresponding to these classes, and then is able to classify an image given its codeword.

Analysis

→ Accuracy Obtained: ~50% (We used 90% of dataset as training data and tested on rest 10%)
→ Kernel used: RBF. Chi2 and Polynomial kernel was experimented with too, but it seemed to be classifying all test samples to one category.
→ Optimal number of clusters: In our experiments, we found 512 yielded good results. We tried with cluster size of 100 and 256 also, but with 512, results were best.

Conclusions

→ Mini Batch K means is faster and gives better results than K means.

→ SIFT is unable to give substantial number of features for images with too many repetitive textures. So, say, an image with a plain background might show no features at all.
→ The accuracy is probably not very good because we are unable to find a kernel that will perform very well on our dataset.