PHOW Image Classification

Juan Camilo Martinez
Universidad de los Andes
Departamento de Ingeniera Industrial
jc.martinez10@uniandes.edu.co

Abstract

A bag of words is a sparse vector of features. In computer vision, these features corrspond to frequently repeating clusters of pixels. However, the BOW method presents difficulty when dealing with spatial data. The Image description by Pyramid of Histograms of Visual Words (PHOW) classification method adresses this limitation

Keywords: bag of words, SIFT, classification

1. Introduction

The Caltech 101 dataset was compiled at the California Institute of Technology in 2003. It consists of 9146 images belonging to 101 visual categories. Each image, aside from being assigned to a single object class, includes a set of annotations with the general bounding box in which the object is located and a detailed human-specified outline enclosing it.

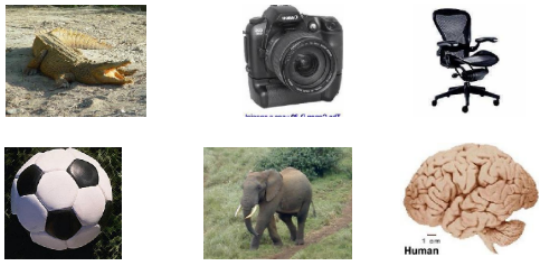

Figure 1. Example of the Caltech 101 Database

The ImageNet dataset is organized according to the WordNet hierarchy, in which each node of the hierarchy represents an object that is depicted by hundreds and thousands of images. Curently, more than 14 billion images have been annotated with the class the objects belong to, and roughly one million also contain bounding boxes.

Figure 2. Example of a very small subset of the Imagenet

The PHOW strategy is an extension of the Bag of Words (BOW) method, that uses Scale Invariant Feature Transform (SIFT) descriptors. The descriptors are then clustered to build visual word vocabularies. Histograms built from the vocabularies in the image are called bags if words. The descriptors used to build the bag of words don't retain spatial information. PHOW adresses this by dividing the image into fine subregions (pyramids) and concatenating the histogram of each of these regions to the histogram of the original while assigning a particular weight to the histograms of the subregions.

Although PHOW includes features extracted with SIFT, it is not scale invariant, as the constructed histogram pyramids depend on image size. While Texton histograms are used as bags of visual words in classifications, they are neither scale invariant nor do they conserve spatial information.

2. Methodology

2.1. vl feat and Caltech 101

The provided example code to implement PHOW presents excellent results for a small sub set of the Caltech 101 Dataset:

The accuracy achieved is of 92%

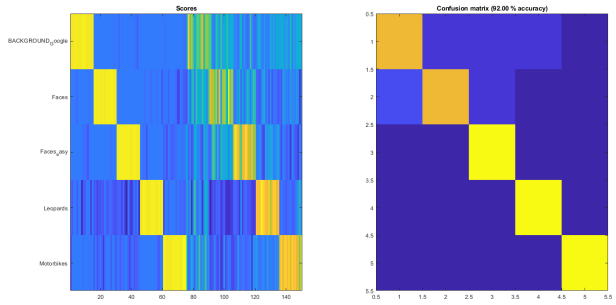

Figure 3. Confusion Matrix for 5 categories of the Caltech 101 dataset

2.2. Results and Evaluation

2.3. vl feat and Imagenet

Using the same hyper parameters for five classes of the Imagenet database produces much poorer results, with an overall accuracy of just over 54%

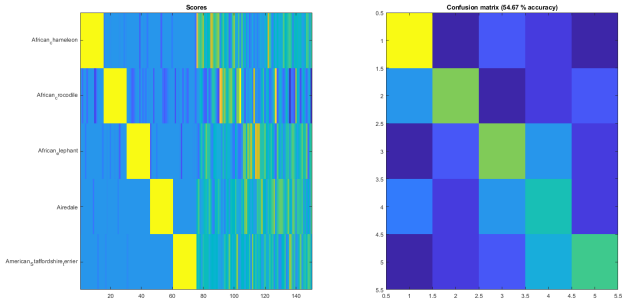

Figure 4. Confusion Matrix for 5 categories of the Imagenet train dataset

Adjusting the amount of names raises the accuracy to 56% but at the expense of classifying some categories more accurately, and confusing others more.

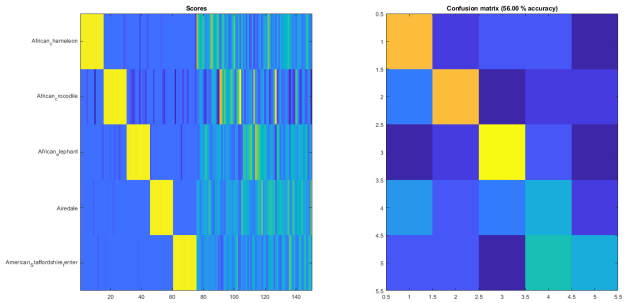

Figure 5. Adjusted hyper parameters

2.4. Classification in Imagenet

Experimentation lead to the discovery that the most relevant yper parameters were number of names (visual words), number of training images and number of PHOW descriotors used to train the bag of visual words (dictionary).

3. Conclusions

.

References