PHOW Lab9

Esteban Galvis
Universidad de los Andes
e.galvis10@uniandes.edu.co

Abstract

The following laboratory classifies images to its respective category by extracting feature descriptors related to visual words in an image (PHOW). The images come from the two popular computer vision datasets Caltech-101 and ImageNet. A Support Vector Machine is used as the classifier.

1. Introduction

Image classification has improved a lot since the famous template matching. The following laboratory uses SIFT descriptors mixed with spatial pyramids and a SVM to classify images to some categories.

2. DataSets

In this lab we will be working with two datasets: Caltech-101 [1] and ImageNet [2]. The first dataset contains 101 categories of images like airplanes, crabs, dolphins, pizza, octopus to name a few. There is a range of 40 to 800 images per category with almost all categories having around 50 images. The size per image has a rough estimate of 300x200 pixels.

The second data set, ImageNet is inspired by Natural Language Processing's WordNet [3] and tries to model 1000 images in average per WordNet synset. There are a total of 100,000 synset in WordNet that correspond to around +80,000 nouns. There are a total of 14,197,122 image and a total of 21,841 synsets. In ImageNet.

3. Method

Recognition has been a main problem in Computer Vision and several methods have been introduced to solve it. The method that this paper will be working with is PHOW [4] which is an extension of a bag of word feature representation created by SIFT descriptors [5].

The SIFT descriptors are built by a scale space of images using Difference of Gaussians at different sigmas. Afterwards, images are searched for local extrema to obtain the best representation of a keypoint at a certain scale. To obtain keypoints that are more accurate, a taylor series expansion is used. Once the local extrema are found, to better represent the keypoint, a histogram of gradients is calculated around it and a feature vector of 128 dimensions is created for each keypoint. The visual words dictionary is created based on these keypoints and are extracted using a k-means algorithm of each 128-feature vector.

PHOW is an extension because it partitions an image into sub-regions and extracts SIFT descriptors in each sub-region creating a spatial pyramid. This extension considers spatial information. However, there is a difference between these two techniques mainly regarding the scale invariant property that SIFT has since PHOW sacrifices the invariance when it partitions the whole image in different pyramids, but compensates with the loss of this information with this same global spatial partitioning. The two main hyperparameters this technique has are L (for how many levels you want to partition) and M (the size of your vocabulary of visual words). All these features are now fed into a SVM classifier.

4. Experiments

4.1. Caltech 101

The baselines we are working with is a tiny problem of the Caltech 101 dataset and the complete dataset. This tiny problem only has 5 categories of the 101 categories that the whole dataset has. The level that is assigned in this problem is 2 and has a vocabulary size of 300. Instead, the complete dataset contains all the 101 categories and has a vocabulary of 600 words since there are more categories, there will be more visual words. The baseline parameters are the best parameters to date since they achieve the state of the art in Lazebnik's paper.

4.2. ImageNet

In this paper we are going to run the dataset with the baseline parameters (level=2, vocab= 600) to see how it performs in this larger and newer dataset.

5. Results

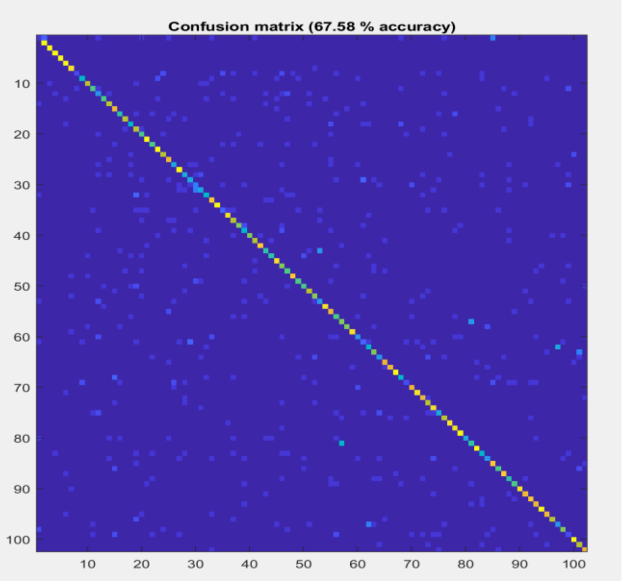

Figure 1: Confusion Matrix for Caltech-101

5.1 Caltech 101

The ACA of the experiment is 67.58%, see Figure 1 which translates to the average classification of all the categories, yet, each class had different results. The classes that were the hardest to classify with PHOW were related to animals like Cougar Body, Beaver, Crocodile and Ant. This is because either the animals didn't have any texture or they could camouflage really well with the environment. The easiest classes were Minaret, Windsor Chair, Joshua Tree, Okapi since some of the presented not a lot of clutter or there were in natural scenes (in which is extremely easy to detect sky, grass and an object).

5.2 ImageNet

Couldn't config the datapath with ImageNet dataset in MatLab.

References

[1] L. Fei-Fei, R. Fergus and P. Perona. One-Shot learning of object categories. IEEE Trans. Pattern Recognition and Machine Intelligence. In press.A. Alpher, , J. P. N. Fotheringham-Smythe, and G. Gamow. Can a machine frobnicate? Journal of Foo, 14(1):234–778, 2004.

[2] A. Berg, J. Deng, and L. Fei-Fei. Large scale visual recognition challenge 2010. www.imagenet.org/challenges. 2010.

[3] Christiane Fellbaum (1998, ed.) WordNet: An Electronic Lexical Database. Cambridge, MA: MIT Press.

[4] Lazebnik, S., Schmid, C., & Ponce, J. (2006). Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 2169–2178).

[5] D. G. Lowe. Distinctive image features from scale-invariant keypoints. IJCV, 60(2):91-110, 2004.