Visual recognition in two different datasets using PHOW

Laura Munar Acosta
Universidad de los Andes
l.munar10@uniandes.edu.co

Maria Ana Ortiz
Universidad de los Andes
ma.ortiz1@uniandes.edu.co

1. Introduction

Object recognition in cluttered real-world scenes requires local image features that are unaffected by nearby clutter or partial occlusion. The features must be at least partially invariant to illumination, 3D projective transforms, and common object variations. On the other hand, the features must also be sufficiently distinctive to identify specific objects among many alternatives. The difficulty of the object recognition problem is due in large part to the lack of success in finding such image features. However, recent research on the use of dense local features has shown that efficient recognition can often be achieved by using local image descriptors sampled at a large number of repeatable locations [5].

Caltech 101 is a dataset with 101 object categories. There are 40 to 800 images per category and most of them have 50 images. The size of each images is in average 300x200 pixels. Most images have little or no clutter. The objects tend to be centered in each image. Most objects are presented in a stereotypical pose. The annotations for all the images consists of a bounding box of the object and a traced silhouette of the objects. This dataset only addresses the challenges of intra-class appearance variation and illumination [3].

ImageNet is an image database organized according to the WordNet hierarchy. WordNet is a lexical database of English in which nouns, verbs, adjectives and adverbs are grouped into synonym sets. These sets are cognitive synonyms that represent different concepts. It contains 200 image classes, a training dataset of 100,000 images, a validation dataset of 10,000 mages, and a test dataset of 10,000 images. All images are of size 64x64 [2].

Using texton histograms for texture recognition is very similar to using bag of feature methods for traditional object recognition. The two methods, texton histograms and bag of features have slightly different inspirations and trajectories. The texton histograms evolved from textons in human vision studies that are approximated by filter outputs and clustered to create textons. The term "bag of features" for visual representation of scenes are inspired by "bag of words" in document retrieval. Both methods cluster image features to create either textons or visual words. The histogram of these textons or visual words is then used to represent the image [1].

Image classification with Pyramid histograms of visual words (PHOW) starts by computing dense Scale Invariant Feature Transform (SIFT). This approach transforms an image into a large collection of local feature vectors, each of which is invariant to image translation, scaling, and rotation, and partially invariant to illumination changes and affine or 3D projection. Previous approaches to local feature generation lacked invariance to scale and were more sensitive to projective distortion and illumination change [5]. After that, the image is partitioned level by level and each level of image is composed of several blocks. Then a series of visual words histograms are formed for the representation of image from low resolution to high resolution in the feature space [4]. The main difference between SIFT and PHOW is that the latter calculates dense SIFT at different resolutions. PHOW is scale invariant because it uses SIFT, which is a scale invariant method to extract features [4]. The purpose of this study is to analyze PHOW behavior with two different datasets: Caltech 101 and ImageNet200.

2. Materials and Methods

PHOW algorithm was used for classification on *ImageNet* and *Caltech 101*. Additionally to test effect of parameters on dataset the methodology consisted on seeing the effect on ACA, the paramaters variated where:

- ***Number of Images in sets***: Train and Test sets where variated to observed the effect on ACA, the amount was varied at the same time.
- ***Number of categories***: It was variated the amount of classes, to test method generalization.
- ***C Parameter***: Variation of slack parameter for svms training.
- ***Number of Words***: Variation of number of words for dictionary.
- ***Num Spatial***: Variation of numspatial for histograms on the method.

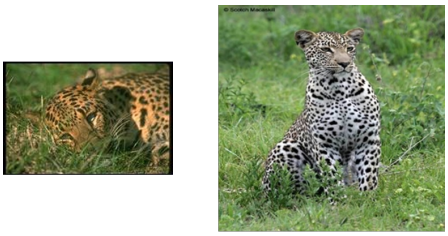

Figure 1. Differences between same category between datasets

Table 1. Performance of Caltech101 dataset with variation in number of classes. The rest of the parameters were not changed

Caltech 101			
# Images	**# Classes**	**ACA**	**Time executed**
30	10	79	38,27
30	15	72,66	64,3574
30	30	67,6	136,5765
30	50	67,33	135,2923
30	100	61,7667	188,5889

Variation of previous parameters where evaluated on Caltec 101 and then used those conclusions for using the method for Image-Net.

Short description of the PHOW strategy (max 2 paragraphs). What is the difference between PHOW and SIFT? Is it PHOW scale invariant? Why/not? What is the difference between PHOW and Textons? Is it worth it? What are the most relevant hyperparameters for the PHOW strategy. Did you found any other relevant parameters inside the script? How can you choose the best set of hyper parameters for Caltech 101 and imagenet set? What are their values?

3. Results

Results obtained are shown in this section, for illustrating difference and difficulty in datasets figures 1 and 2. Results from variating number of words are show on Table 2 and for variating number of classes are shown on Table 1. Aditionally on Figure 3 it is possible to observed the change on number of images versus accuracy score, as it is observed as a decreasing line. In Figure 3 the effect of ***C*** parameter on Caltech 101 can be observed on Figure 4, where no particular pattern can be observed, however some peaks on 300 and 600 words are very similar. Finally on figure 6 the confusion matrix over the training set obtained for best hyperparameters can be observed and it is possible to observed how for categories 30-40 color isn't as bright.

Evaluate the performance (ACA) of the classifier in the train set of Image-Net and Caltech 101 and imagenet. Report a single number for the whole set. Why is there Csuch a big difference?

Figure 2. Difficulty of recognition within datasets

Table 2. Performance of Caltech101 dataset with variation in number of words. The rest of the parameters were not changed

Caltech 101			
# Images	**# Words**	**ACA**	**Time executed**
30	50	61	96,9774
30	100	64,933	82,6158
30	150	65,8667	85,5655
30	250	66,2667	98,4155
30	300	67,2	97,4549

Table 3. Results of variating num spatial for PHOW method on Caltec101

Caltech 101			
# Images	**NumSpatial**	**ACA**	**Time executed**
15	[2,4]	67.58	566.34
15	[4,2]	68.10	1065.86
15	[4,8]	66.21	866.52
15	[9,4]	65.55	923.27

4. Discussion

Performance (ACA) Using your classification function on imagenet test, again a single figure. What seem to be the easy classes, what seem to be the hardest? Why? Show pictures, analyze model behavior. Not stay just with "The easyest are cars, the hardest flowers." Yeah, and? Elaborate. .

The most important hyperparameters for the PHOW algorithm are: number of classes, number of dictionary words, number of images in train and the C parameter of the SVM. There are other parameters than can be considered important and as a future work we recommend to use are the Color Space of the images, the size of the Gaussian window used to smooth the image and the kernel in SVM classifier. According to the table 1, when only considering 10 classes the PHOW algorithm has its best performance (ACA 79). This can be explained with the fact that SVM is a binary classifier that can be adapted to multiple classes but this pays off on the performance of the algorithm. So when there are less classes (10 vs 101) the classifier works better.

On the other side, the number of words in the dictionary is equivalent to the k parameter in k-means. Meaning

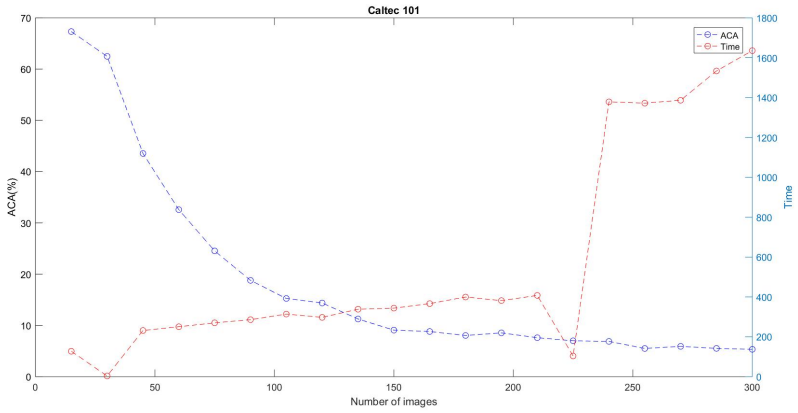

Figure 3. Varying number of train and test images effect

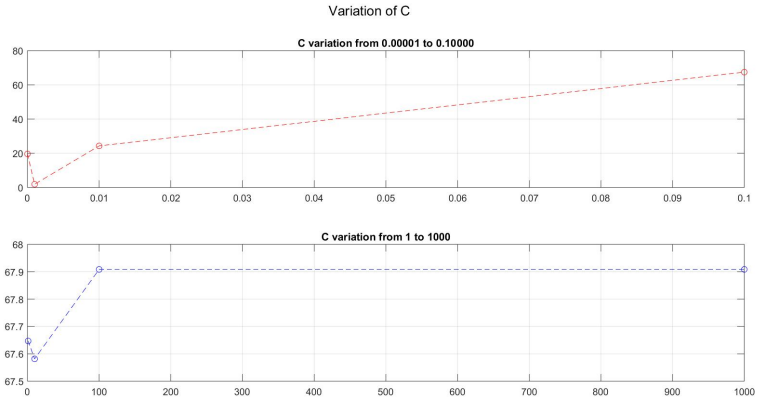

Figure 4. Varying slack parameter on SVM's

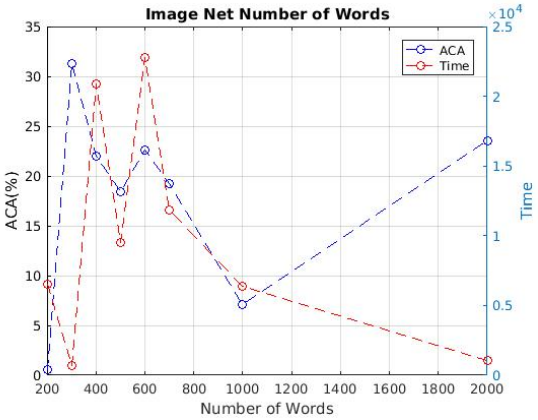

Figure 5. Results obtained for variating the number of words for image net

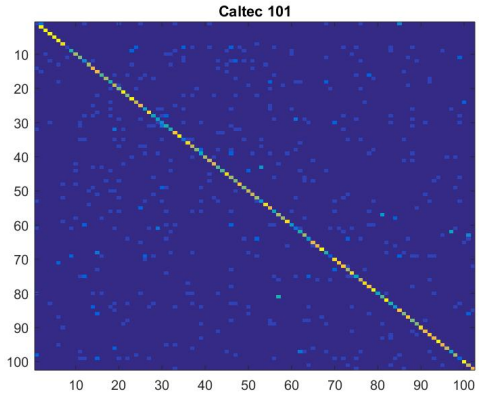

Figure 6. Confusion Matrix obtained from best method

that it has to be determined in an iterative way. The table 2 presents the experiments performed in the Caltech101 dataset for a variety of number of words in the dictionary. Here it can be seen, that greater number of words leads to a better performance of the algorithm. According to what it is stated in literature, few words results in a small dictionary that is not representative for the dataset. Otherwise, too many words can lead to overfitting the training set.

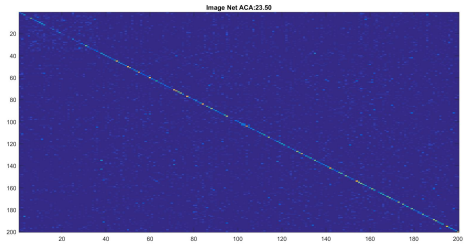

Figure 7. Confusion Matrix obtained from best method on Image Net

It has been pointed out that the categories that have more pictures are easier to classify (Airplanes, Motorcycles, Faces). That can be explained because the classifier has a better description of the category with a higher number of images.

On Figure 3 it is observed how Accuracy scores lowers its value as number of images on sets growths, this effect can be address as model is not generalizing enough and how model has issues on giving enough enough words for each class, revolving in issues, this trend may be fixed by adding more words as training and test set growths. Also this means model is not characterizing enough variability in each class. Additionally,on Figure 4 it can be observed how the effect of different slack values on SVM may affect accuracy score, and it is clear how values higher than 100 have a constant behaviour and lowers have a increasing line behaviour, however values lower that 0.1 have a poor development. This is caused because more slack error is higher allowing less strick barriers, and possibility to allow variabilty within the classes.

Even though, Caltech101 dataset states that most images have little or no clutter, the objects tend to be centered in each image and most objects are presented in a stereotypical pose. It can be seen in Figure 2 the difficulties of extracting features when the object is hidden or in the image are only parts of it. Also, Figure 1 compares the category Leopards between the datasets in study.

The main challenges of the recognition in the both datasets are that, for humans is obvious to recognize and classify objects even though these objects are translated or only a part is present in the image; but this task is not obvious for a computer. So, it is important to find the features that best describe each category in order to give as much information as it can be to the classifier. On the other hand, in order to improve the results a change of classifier is highly recommended. SVM are great binary classifier but in this problem we are talking about over 100+ classes. As future work, Random Forest classifier should be implemented. For caltech 101 confusion matrix on Figure 6 it can be observed that some categories are not corrrectly classify, mainly because its categories are not describe enough. Additionally for Image-Net on **??**

For Image-Net on figure 5 the effect on accuracy score on variating the number of words it's very arbitrary, best were 200, 600 and 2000 number of words, being the highest ACA 32%, which is very low comparing values obtained from Caltech101.

Method result variation on both sets appear mainly because sets are very different, not only on image quality, but on variation of images. Additionally that method was develop over caltech101 and other parameters that were not evaluated may have been very important.

5. Conclusions

PHOW method does not work correctly on Image-Net, as mentioned before becaused of variation on images. Additionally numwords for image net is not an accountable parameter and does not has any particular parttern. Parameter that most affected accuracy score for caltech101 was the number of classes, however a simple version of the problem was not the main strategy, that's why most important parameters was the amount of words and the amount of images in sets, which is a problem because of the particular performance of this algorithm, were more information results on lower ACA.

Improving this project can be done by using Random Forests because is a multiclass algorithm and is computacionaly more efficcient.

References

[1] K. J. Dana. Computational texture and patterns: From textons to deep learning. *Synthesis Lectures on Computer Vision*, 8(3):1–113, 2018.

[2] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[3] L. Fei-Fei, R. Fergus, and P. Perona. One-shot learning of object categories. *IEEE transactions on pattern analysis and machine intelligence*, 28(4):594–611, 2006.

[4] H. Gao, W. Chen, and L. Dou. Image classification based on support vector machine and the fusion of complementary features. *arXiv preprint arXiv:1511.01706*, 2015.

[5] D. G. Lowe et al. Object recognition from local scale-invariant features. In *iccv*, volume 99, pages 1150–1157, 1999.