Image Classification on Caltech 101 and ImageNet with PHOW

Laura Bravo
Universidad de los Andes
Cra 1 #18a-12, Bogotá, Colombia
lm.bravo10@uniandes.edu.co

Juan Pérez
Universidad de los Andes
Cra 1 #18a-12, Bogotá, Colombia
jc.perez13@uniandes.edu.co

Abstract

Object recognition/classification from images is one of the fundamental tasks in Computer Vision. Regarding the present work, the task consisted on taking an image, depicting usually a single large object whose class is known, and assigning a class label to it; naturally, this requires a dataset with these specifications, so two well known datasets were selected: Caltech 101 and ImageNet. Caltech 101 is a rather small dataset that was used in the middle 2000's, while ImageNet is a large and newer dataset that is currently used for a variety of tasks in Computer Vision (given its size, only a subset was taken). For performing this task, the strategy known as "Pyramid histograms of visual words" or "PHOW" was used. This approach computes Scale Invariant Feature Transform or SIFT descriptors and interprets them as visual words, from which a histogram of words of a given image can be computed and used to classify the image. The actual implementation of this strategy that was used was the VLFeat library, which is open source and written by A. Vedaldi and B. Fulkerson. Several hyperparameters of this method were varied independently when training the algorithm over both datasets, looking to improve its performance over the test sets and, finally, an overall accuracy of 73.82% over Caltech 101 was obtained, while this figure was 27.99% for ImageNet.

1. Introduction

Classification of images consists on assigning a label to an image that reflects its contents, usually a certain type of object, for instance, a dog. It is one of the fundamental problems in Computer Vision, since it is the basis for *understanding* an environment once individual objects have been detected. It has been widely studied by many research groups across the world, and very important advances have been made in the last years [1].

As many problems in Computer Vision, the approach taken is to select an image in which some object appears, extract some features that characterize the image and compare it to previously known features of other images whose class is known. Then, according to this information, a label is assigned to the image.

A strategy for doing this is Pyramid Histograms of visual Words (PHOW), which uses several concepts, like SIFT, visual words and histograms of those visual words. SIFT is probably the most popular feature used in Computer Vision, since it detects salient image regions (keypoints) and extracts discriminative yet compact descriptions of their appearance (descriptors). Its keypoints are invariant to viewpoint changes that induce translation, rotation, and rescaling of the image [2].

The idea is to take these features from a dataset of images and create a visual dictionary, that is, to look for some basic elements (*visual* words) that, when taken together, form images. With this vocabulary at hand, an image can be described as a histogram showing the frequency of appearance of every visual word in the dictionary in it. What would be expected, then, is that images containing similar objects have similar histogram of words, and, consequently, if the class of some images is known, new images could be classified based on that.

In the present work, PHOW was used in Caltech 101 and a subset of ImageNet. A simple attempt to optimize its hyperparameters according to its performance on the *train* set of each dataset was made and, when these "best" parameters were found, they were given to the algorithm for the resulting method to be evaluated over the *test* set.

2. Materials and methods

2.1. Datasets

Throughout these experiments, two datasets were used: Caltech 101 and ImageNet. From each of these, models were trained in the *train* set and, then, were evaluated in the *test* set. Particularly, due to Caltech not having this intrinsic dataset division, it was created when as a parameter of the method. The implementation of PHOW used was that provided by the VLFeat Library written by A. Vedaldi and B. Fulkerson [2].

Caltech 101 is a dataset of pictures of objects belonging to 101 categories with about 40 to 800 images per category (most categories have about 50 images). It was collected in September 2003 by Fei-Fei Li, Marco Andreetto, and Marc 'Aurelio Ranzato. The size of each image is roughly 300 x 200 pixels [3]. It is a much smaller dataset than ImageNet, but still it can be used to compare the output of some methods, and provide a preliminary result.

ImageNet is an image dataset organized according to the WordNet hierarchy. Here, each meaningful concept in WordNet, possibly described by multiple words or word phrases, is called a "synonym set" or "synset". There are more than 100,000 synsets in WordNet, from which the majority are nouns [4]. Images in this dataset are not 'ideal' images of objects but, rather, actual images of how objects look in real life, showing occlusion, different views, intra-class diversity, etc. Given that this is an enormous dataset, a subset was taken, in which there were 200 classes and 100 instances per class.

2.2. PHOW

In order to describe and classify the images in both datasets, Caltech 101 and ImageNet, the PHOW strategy was used. It consists on using a Bag of Words approach (BoW), that is, to use telltale keypoints in images and use them to generate a descriptor based on their frequency of occurrence. This was accomplished by executing a dense Scale Invariant Feature Transform (SIFT) on each of the training images and clustering the resulting descriptors in order to obtain the visual words, namely, in each pixel of an image a 128 dimension descriptor was extracted, which was the result of the convolution with 8 orientation filters in a patch around the pixel composed of 16 cells and the corresponding orientation histogram for the patch with 128 bins. The clustering was performed with K-means, with the centroids representing the visual words.

Afterwards, each image in the training set is characterized based on the frequency or histogram of the visual words. To include spatial information, the PHOW strategy utilizes a pyramidal representation, where the concatenation of the word histograms in each level is the final descriptor. An SVM is later trained to classify the images, in reference to the final word histogram descriptor. Nonetheless, due to the binary nature of SVMs, it is necessary train one SVM per category of interest. To classify an unknown image, its pyramid of word histograms is calculated and then this is fed to the SVMs, where the category of the image will be decided by the SVM with the highest confidence. To improve the speed of PHOW, decision trees may be used so as to reduce the number of required comparison for assigning a word to a given patch.

When using PHOW, there are multiple parameters that must be accounted for, among the most relevant are: the number of training images, they determine the ability of the method to understand intra-class variance, for this reason it is highly related to the number of test images; the number of words, this is decisive for avoiding over and under fitting; the cost parameter, kernel and solver of the SVM; and the spatial partitioning used in the pyramidal representation, that is the dimensions of the patches and the levels of the pyramid. To obtain the best result for this strategy, these were varied in both datasets (Caltech 101 and ImageNet).

The evaluation methodology consisted on calculating a confusion matrix per dataset. It depicts the results of the classification, in terms of the categories where the images of the test set where placed. For each category, it quantifies the number of images in said category that were correctly placed (according to the ground truth) and those incorrectly classified. The mistakes are distributed in the categories that the method predicted. Therefore, the success of the algorithm corresponds to the number of images situated in the diagonal of the matrix, given that its rows are the ground truth categories and columns the predictions. To solve the issue of a miss balance in the number of images per category, then, the matrix is normalized according to the number of instances in the ground truth. Furthermore, the mean of the main diagonal in the normalized matrix is known as the Average Classification Accuracy (ACA). This will be the ultimate result of the evaluation.

2.3. Estimation of hyperparameters

The estimation of hyperparameters for a given model is a fairly complex problem of optimization with quite a large number of variables to take into account. Additionally it has a very important constraint in this case: computational power. For instance, each time the algorithm is trained with some given parameters in the *train* set of ImageNet (taking only 40 images per class), it takes about 1.5 hours to run on a server; this clearly limits the amount of experiments that can be run within a given timeframe.

The approach here was to vary parameters independently, train the models in the *train* set and measure their performance on the *test* set. Then, some claim about the behavior of the system could be made and each parameter could be chosen on its own, so that a model containing the parameters that had the best performance could be run. This is a mere approximation to finding an optimum of the system, at least in a given range, because this is making the enormous assumption that there is no interaction between each of the parameters, i.e. the system is linear, which clearly is not true in this case. Nevertheless, if the range in which the parameters are chosen is not too big, it could be an acceptable guess.

Given that the two datasets are very distinct in size, and therefore the training times differ quite a bit, different experiments for the hyperparameters were run for each

dataset. Overall, the parameters that were varied were: soft margin (C) for each SVM, number of words and spatial parameter (scales for the pyramids).

3. Results

3.1. PHOW over the datasets

In ImageNet, there are fined grained classes, the differences among them are slight, meaning that two categories may share a large part of the visual vocabulary that describes them. Thus, the algorithm lacks the resolution needed to distinguish them. This is aided by calculating dense SIFT on the whole image, as opposed to using keypoints. Theoretically, the spatial invariance of SIFT should respond well to view point differences and acceptably towards large enough, only partly occluded objects. But, it falters when faced with a changing background (clutter and the position of the object). This is due to the noise that the background adds to the representation. Whereas in Caltech 101 the vocabulary describing the images was sufficiently consistent to favor classification, in ImageNet the vocabulary extracted from the whole image may not represent the object of interest but the scene. Also, the scenes are not consistent for a category, consequently, the descriptor fails to take the object as the relevant aspect in the image producing misclassification.

3.2. Experiments on Caltech 101

After sufficiently varying the hyperparameters of PHOW in Caltech 101, the best ACA obtained was 73.82%, this corresponded to using 1000 words in the visual dictionary; a C of 10 and the *svga* solver for the SVMs; 30 images for training and a symmetrical spatial pyramid whose levels were [2 5].

In order to obtain this result, several experiments were performed. Initially, given a fixed number of words, 400, the parameter of the size of the individual patches in the spatial pyramid was varied. This was trained on 30 images and tested on 15 or the remaining images. The slight effect on the ACA is shown in Fig. 1. The best result, 67.58%, was obtained by using two levels in the spatial pyramid, the first with 2 and the second with 5 pixels as the patch sizes.

Nonetheless, the change is not sufficiently significant to unequivocally state that this is the best pyramid configuration. For this reason, a second experiment shown in Fig. 2 used a fixed size of the spatial pyramid [2 5] and with this the number of words was varied. A more prominent change in the ACA can be observed. The graph suggests that increasing the number of words has a directly proportional effect on the ACA. However, the cusp of the number of words was not reached (the optimum), due to computational requirements, but the slope is beginning to stabilize.

Once the number of words was established as 1000, the

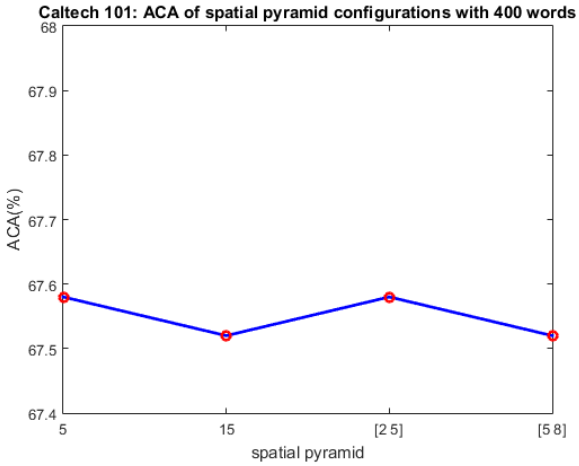

Figure 1. Accuracy obtained by PHOW on Caltech 101 with various values for the spatial pyramid parameter. There were 400 words in the vocabulary. The best was obtained with [2 5] and [5]. The ACA was 67.58%

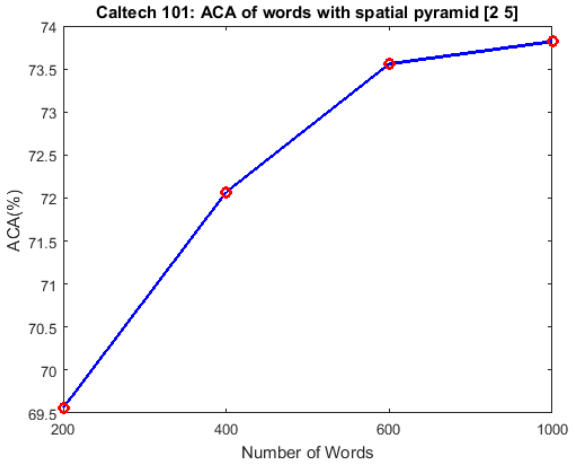

Figure 2. Accuracy obtained by PHOW on Caltech 101 with various values for the parameter number of words. The spatial pyramid was [2 5]. The best was obtained with 1000 words. The ACA was 73.82%

size of the spatial pyramid was varied once more. Fig. 3 depicts the effect of varying this parameter on the ACA. Increasing the complexity of the pyramid was inversely proportional to the method's accuracy. A possible explanation could be that while there is an increase in quantity of information, it may not provide the representation with additional data. In fact, it may introduce more noise by partitioning the desired object into segments not compatible with the visual vocabulary.

The result of the compendium of the best parameters, 1000 words, 30 training images, an SVM cost C of 10 and a spatial pyramid of [2 5], can be found in Fig. 4. The normalized confusion matrix shows an almost random error in

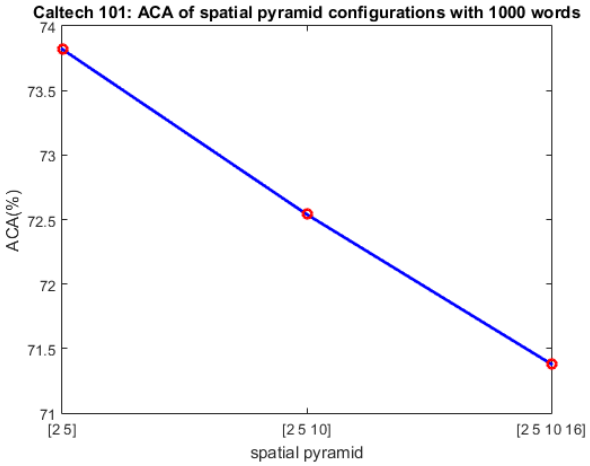

Figure 3. Accuracy obtained by PHOW on Caltech 101 with various levels for the spatial pyramid parameter. There were 1000 words in the vocabulary. The best was obtained with [2 5]. The ACA was 73.82%

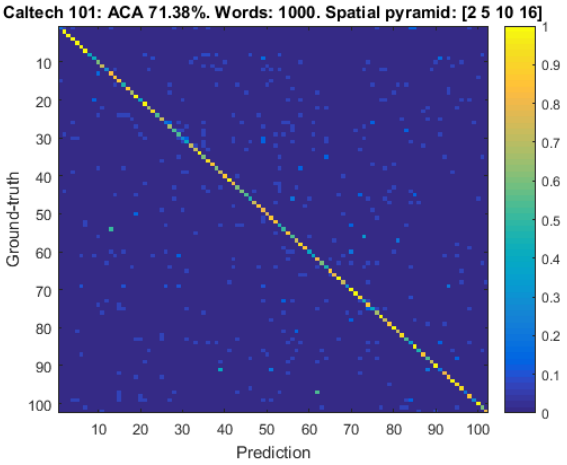

Figure 4. Confusion Matrix obtained by PHOW on Caltech 101 with 30 training images, 1000 words, SVM cost C of 10 and a spatial pyramid of [2 5]. The ACA was 73.82%

Classes	Accuracy (%)
Stop Signs	100
Trilobites	100
Snoopy	100
Saxophone	100
Background Google	0.2
Crayfish	0.2
Waterlily	0.2
Strawberry	0.2

Table 1. Best and worst performing classes on Caltech 101.

the classification, that is, there is no obvious source of confusion for the algorithm. Nevertheless, upon closer examination, the worst categories (1) such as Background google and Crayfish have highly different backgrounds. This variability could be adding a large amount of noise to the descriptor, muddling the choice of visual words and thus the method's accuracy. A possible alternative for decreasing this source of error could be having an optimization function between the expected keypoints or visual words of the image and the one actually extracted. Thus, iteratively improving the descriptor on subsets of the training images but using supervision. On the other hand, in the best performing or easier categories 1, there were several where the accuracy was perfect (1), all have distinctive features that make it easy to determine shape, particularly the stop signs. For instance, trilobites are all in the same orientation (with images having black triangles) and the fossil is cut in a characteristic oval shape that simplifies the task. The effect of the shape is also apparent when examining the saxophone images. Because they are mostly profile images, there is a repetitive "L" shape. This behavior is also seen in the Snoopy category, being a cartoon he is consistently the same abstraction of a Beagle dog.

The results obtained on Caltech 101 and the subset ImageNet differ significantly despite using the same approach (PHOW). The main reason for this is the nature of the datasets themselves. Whereas Caltech 101 focuses on intra-class variation, disregarding other challenging aspects of classification, ImageNet embraces these challenges incorporating them fully. While the images from Caltech 101 are focused on the object, usually with a simple background and of roughly the same size, the images in ImageNet include the intra-class variation, but also it varies highly in the object position, the view points, background clutter and occlusions [3, 4].

3.3. Experiments on ImageNet

The performance of the classifier in the *train* set of ImageNet was found to be 21.1625% with the default parameters, that is: 600 words, [2 4] as the spatial parameters (equal for x and y) and 10 as the soft margin (C) for the SVMs. Tremendously low when compared to the performance with those same parameters over Caltech 101.

All the subsequent experiments were performed by preserving these default parameters and only varying one of them at a time, so that a linear approximation of the system was being assumed.

The C parameter was varied between 5 and 25, with steps of 5, and the results are shown in Fig. 5. Apparently, with these numbers, there is not much of a change (less than 0.2% points were increased). Given that no other experiments were run due to time, the best parameter from this experiment was taken to be 25.

The spatial parameters in the algorithm, that control which and how many scales are taken into account when

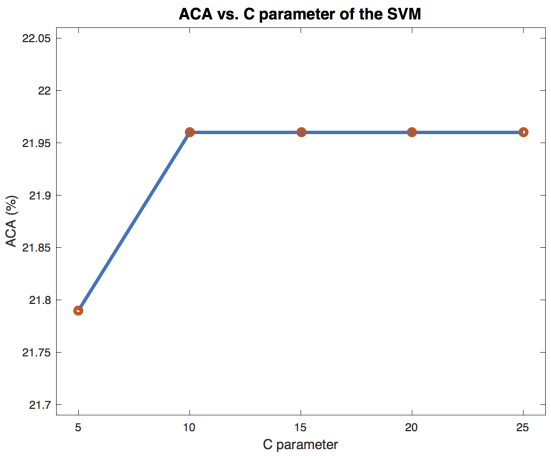

Figure 5. Accaccuracy obtained by PHOW over the *test* set of ImageNet with various C parameters for the SVM.

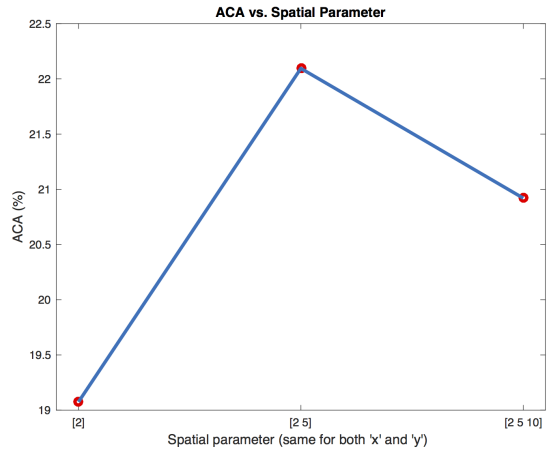

Figure 6. Accuracy obtained by PHOW over the *test* set of ImageNet with various spatial parameters. Note: the parameter was taken to be equal for both x and y directions.

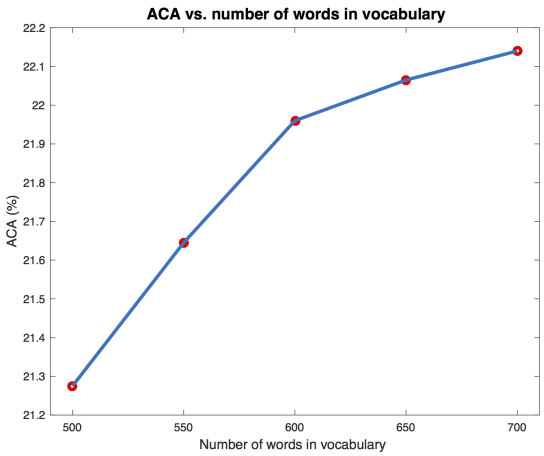

Figure 7. Accuracy obtained by PHOW over the *test* set of ImageNet with various number of words in the vocabulary.

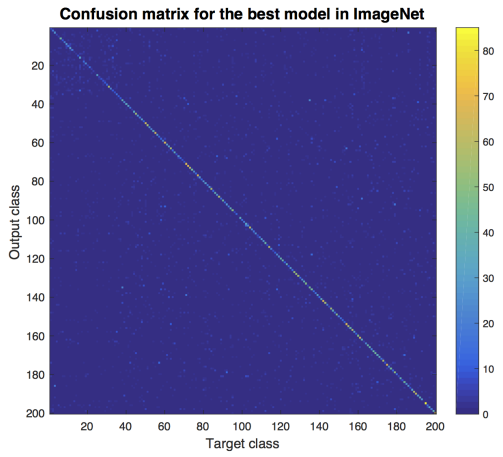

Figure 8. Graphical display of the confusion matrix of the output of the best model over the ImageNet dataset.

learning and testing, were varied in an increasing manner: first only one number was used, 2, then two numbers were used, [2 5], and finally three, [2 5 10]. Even though experiments with [2 5 10 16] were run multiple times, turns out that this floods the server's memory when training, making it impossible to obtain a model.

The results of this experiment can be seen in Fig. 6. Some strange behavior is observed, it appears to have a maximum in the center parameter [2 5], which is, at first, counterintuitive, since one would think that as the number of scales increases, so does the amount of information and, consequently, choices made by the algorithm should improve. However, this showed that the best parameter, within the ones that were tried, was [2 5], therefore it was chosen to be the winner of this experiment.

An experiment inspecting the number of words in the vocabulary that was created was executed. Number of words varied from 500 to 700 with steps of 50. Results are displayed in Fig. 7. For the values in which this was varied, an asymptotic behavior is observed: each time the number of words is increased by 50, the algorithm improves by less, so that it could be argued that it will tend to some value when the number of words increases beyond 700. Nevertheless, this was not further explored due to time resources, and 700 was taken to be the parameter that had the best performance.

Finally, two models with the best parameters that were found, 25 as the C parameter of the SVM, [2 5] as the spatial parameter for both x and y and 700 as the number of words in the vocabulary, were trained with (i) 40 images of the train set, which is the number of images that all the experiments were trained with, and (ii) 100 images of the train set.

Then, both models were evaluated over the *test* set of Im-

ageNet. The first one had an ACA of 22.395%, while that number for the second model was 27.990%. This performance is clearly higher than that gotten for any independent parameter tuning, which supports the idea that the combination of the best parameters actually provides a better overall accuracy for the model. The confusion matrix for this "best model" on ImageNet is shown in Fig. 8. For this model, the four easiest and hardest classes in which to classify images are shown in Table 2.

Classes	Accuracy (%)
Website	85
Zebra	74
Coral fungus	68
Carbonara	67
English setter	0
Weasel	0
Italian greyhound	2
Hog	2

Table 2. Best and worst performing classes on ImageNet.

With the aim of determining why these are the easiest and hardest classes, a visual exploration of the first image of each of these classes is shown in Fig. 9. Here, some reasons as to why this is can be stated:

- In general, animals are hard to classify (except for the zebra, which is a very particular animal that shows unique texture). This is probably due to the fact that animals share specific shape: face with all the usual elements, snout, not more that four legs, usually black eyes.
- Regarding the previous statement, it could be argued that if there was a super class containing animals instead of specific animals, the performance of this method would probably increase significantly.
- Websites have very unique features that make them very different from all the other classes. For instance, everything is very well organized and structured, there are probably letters, and elements tend to be on the same places.
- All the best performing classes show unique shapes that do not resemble other classes in ImageNet. The coral fungus, for instance, has strong texture in a very characteristic pattern that does not change much with the scale at which it is processed.

From these observations one very important lack of the performance evaluation is made clear: no information regarding taxonomy of the classes of these objects is being taken into account. Object that are closed taxonomically,

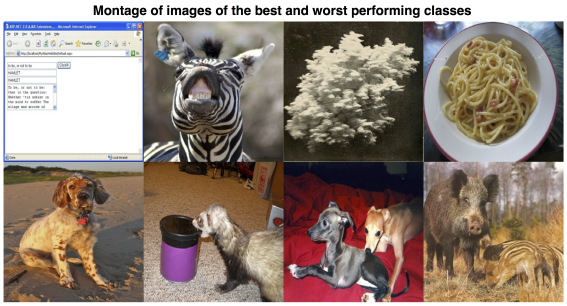

Figure 9. Montage depicting images of the best and worst performing classes on ImageNet. First row are the best performing classes while the second row are the worst performing classes.

in WordNet, have a very high probability of looking alike, and therefore two things should be made: (i) development of a classification model that knows about this and, when in doubt about differentiating objects that are taxonomically close, chooses to give the class that contains them as the output, therefore increasing its accuracy (or, for instance, it could give the taxonomy from which it is choosing), and (ii) development of a metric for evaluating performance that takes it into account, since it is not the same confusing a Golden Retriever with a German Shepherd as confusing a Golden Retriever with a toilet; naturally, this metric should consider the distance in the dendrogram, and not only "hit" or "not hit".

4. Conclusions

The big difference in the performance of PHOW on both datasets can be explained by difference in the difficulty of the images in each dataset. ImageNet proposes images that are much more "natural", implying that there are more occlusion, deformation, intra-class variation, differences in brightness, etc.Apparently, from the results, having two spatial scales is better than just having one, but more than this decreases the method's accuracy.

Dictionaries with a large number of words, over 600, tend to perform better than those below this threshold. Nevertheless, it is quite possible that this depends entirely on the dataset, considering that as the dataset gets larger it is more likely that new visual words will appear, increasing the dictionary's size. Additionally, the PHOW approach presents many difficulties besides computational restrictions, in general they are due to the chosen visual words not being able to accurately describe the images on the datasets. This was due to the intrinsic restrictions of the descriptor, but also because of the noise added by the background and the sampling of the image. However, the method performs acceptably well on datasets with little variation in the images other than the categories, such as Caltech 101. But it fails to generalize when faced with a more varied dataset analogous to ImageNet. Finally, the very different results obtained

on both datasets are a clear example of the scope of this method, but also of the highly dataset dependent and slow process that is adjusting parameters. The ones found for Caltech 101 could not be directly translated into ImageNet.

References

[1] "UC Berkeley Computer Vision Group - Recognition", Www2.eecs.berkeley.edu, 2017. [Online]. Available: https://www2.eecs.berkeley.edu/Research/Projects/CS/vision/shape/. [Accessed: 06- Apr- 2017].

[2] A. Vedaldi and B. Fulkerson, "VLFeat - An open and portable library of computer vision algorithms", pp. 1-4, 2010.

[3] "Caltech101", Vision.caltech.edu, 2017. [Online]. Available: https://www.vision.caltech.edu/Image_Datasets/Caltech101/#Description. [Accessed: 04- Apr- 2017].

[4] "ImageNet", Image-net.org, 2017. [Online]. Available: http://image-net.org/about-overview. [Accessed: 04- Apr- 2017].