PHOW classification over Caltech 101 and ImageNet

Stephannie Jimenez Gacha
Universidad de los Andes
Cra 1 Este No 19A - 40 Bogotá, Colombia
s.jimenez16@uniandes.edu.co

Sergio Galindo León
Universidad de los Andes
Cra 1 Este No 19A - 40 Bogotá, Colombia
sa.galindo10@uniandes.edu.co

Abstract

Pyramids of Visual Words (PHOW) are an extension of the bag of words with the special feature that spacial information is included. The performance of Caltech PHOW implementation was measured using not only the Caltech101 database but also with a tiny subset of ImageNet. The default parameters were tested and compared versus some modifications to see the effect in both datasets. The maximum ACA registered was of 70.8% in Caltech101 and 19.1% in ImageNet. The huge difference of the percentage of classification between the two is due to test and train subsets of Caltech101 are very similar. This means, that the variability is not elevated determining that Caltech101 is an easier database. As well, the use of spatial information increases as the size of the cell grid is smaller.

1. Introduction

Caltech 101 and ImageNet are two of the most important classification datasets in computer vision. The first one, introduce by Fei-Fei in 2003, is composed of 101 classes with 40 to 800 images per class[1] while the second is a very large dataset composed of nearly 15 million images classified into 22000 classes[5]. These two datasets have been used extensively in the recognition tasks since their introduction and, specially ImageNet, has become a common benchmark for classification problems. While Caltech 101 has the disadvantage of centered images with objects in stereotypical poses and little clutter, ImageNet presents an outstanding variability for each class, allowing a better approach to real images and better performance in practical applications.

One of the common representation used in image recognition is the pyramid of histogram of visual words (PHOW), an extension of the bag of words that incorporate the spacial information. In this method, the feature invariant scale transform (SIFT) is used to obtain and represent the main images features. After that, a visual words dictionary is constructed by clustering of the features into k centroids(or words), such that, the images can be represented using the concatenation of their histogram of visual words at different scales[2].

Then, in order to accomplish image recognition, a classifier(usually SVM) is trained in the feature space. To classify an incoming image the PHOW representation of obtained and a distance metric is applied to label the incoming data into any of the classes. Commonly, the SIFT descriptor is a 128 dimensional vector containing the gradient orientations of the patches. This which implies high computational resources for clustering, also, the number of words is critical in the correct classification as the vocabulary should be sufficient enough without to much redundancy[3].

For this, an evaluation of the PHOW methodology using the VLFeat library over the Caltech 101 and a subset of the ImageNet dataset is proposed. The results over both datasets will be compared using the average classification accuracy and the confusion matrices, and the optimal parameters for each dataset will be used to determine the optimal hyper parameters for each database.

2. Materials and Methods

For both datasets, PHOW was used for image representation and a SVM was trained as a classifier. The phowcaltech101 function of the VLFeat library was used in all cases and the ACA as well as the confusion matrices were obtained. Due to the large size of the ImageNet database it was subsample to a total of 200 classes.

2.1. PHOW representation

To obtain the histogram of visual word representation of the images the phowcaltech101 function with the default parameters was used. In this configuration the function selected 15 random train images from each class to obtain the train features, which means, for caltech101 the dictionary of visual words was constructed from 1515(100 classes) images and for ImageNet from 3000 (200 classes). After that, the function computes the SIFT descriptor for each patch of interest producing a 128 dimensional vector for each one.

After setting the number of words to 600, the respective codevectors were obtained by k-Means. In addition, the quantization of features was performed using a k-dimensional tree that makes n(=600) partitions in the feature space to find the visual words.The final histogram representation of each image was obtained by means of concatenated spacial histograms of visual words at 3 different levels of the spacial pyramid.

2.2. SVM train and evaluation

To classify the images a SVM was trained using the PHOW representation of the train images. The C parameter of the support vector machine was fixed to 10, corresponding to a small margin hyperplane[4]. The function by default selected 15 random test images from each class and the label of each image was determined by stochastic dual coordinate Ascent. The ACA and confusion matrix for each dataset was obtained.

2.3. Experiments

In order to asses the influence of the train and test number of images, the number of classes, visual words, number of windows in the spacial pyramid and the C parameter of the SVM we performed the classification over the Caltech 101 dataset varying each of the parameters one at a time. The ACA of each experiment was registered.

3. Results

We obtained results for both databases, Caltech 101 and ImageNet. In Caltech 101, seven experiments were made. The main idea was to deeply understand the function of *phow_caltech101.m* and the effect of each parameter of interest in the test set. In this way, the results in the test subset of ImageNet can be improved. As it is a classification task, we calculated the confusion matrices for each run.

3.1. Caltech 101

In the first place, the default configurations of the script were tested and the results are shown in figure 1. The *default* parameters were with the complete classes, with 15 images of train, 15 images of test, 600 visual words, the spatial X and Y in [24] and the plane hyperparameter for the SVM, C in 10.

In order to measure the effect of each parameter previously mentioned, the default settings are maintained and the only change is for the parameter. The figures from 2 to 7 contain the results of each experiment.

The table 1 helps to visualize the ACA results for each one of the trials in an effective form.

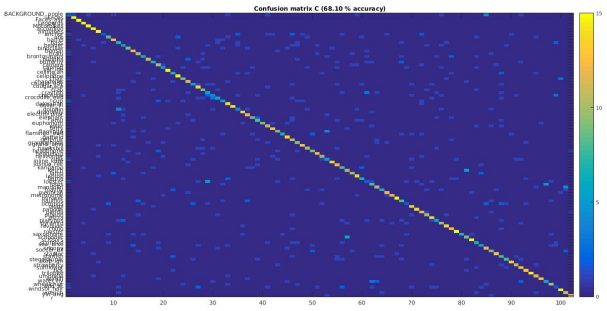

Figure 1. Confusion Matrix with default parameters in Caltech 101.

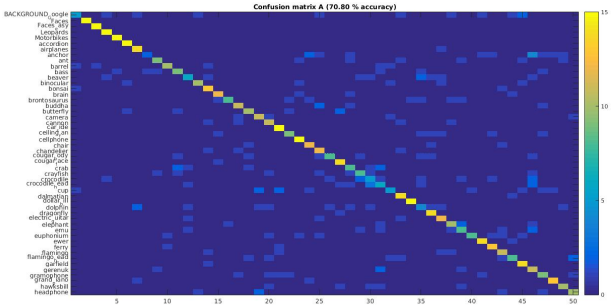

Figure 2. Confusion Matrix for 50 classes

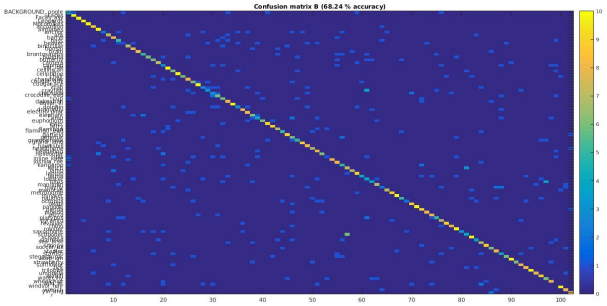

Figure 3. Confusion Matrix with 10 test images and 15 train images.

Table 1. Results of ACA in Caltech 101 database

Parameters	**ACA**
Default	68.1%
50 Classes	70.8%
10 test, 15 train	68.24%
20 test, 10 train	61.96%
300 visual words	66.99%
C = 5	68.1%
X,Y = [4 8]	66.01%

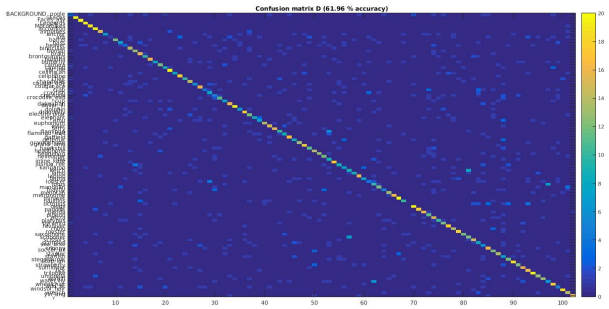

Figure 4. Confusion Matrix with 20 test images and 10 train images.

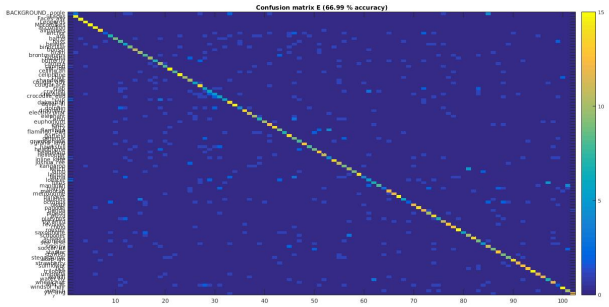

Figure 5. Confusion Matrix for 300 visual words

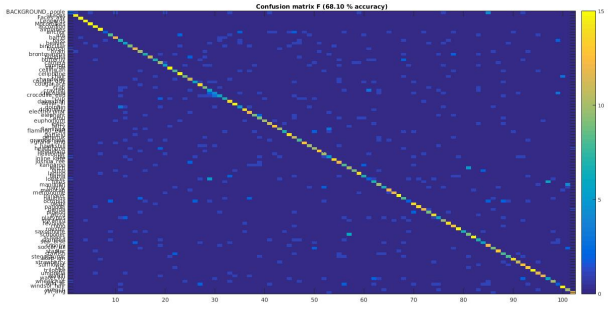

Figure 6. Confusion Matrix changing the SVM hyperparameter C to 5.

3.2. ImageNet

In the first place, this database was tested using the default configurations already explained. The results were not as high as expected, the ACA is 15.7% and figure 8 shows its confusion matrix.

Using the information obtained from caltech101 database, a reduction of classes was attempted with a major number of training images than test images. This was the selected model in order to run the test of ImageNet. With this configuration the ACA obtained is 19.4% and figure 10 shows the confusion matrix.

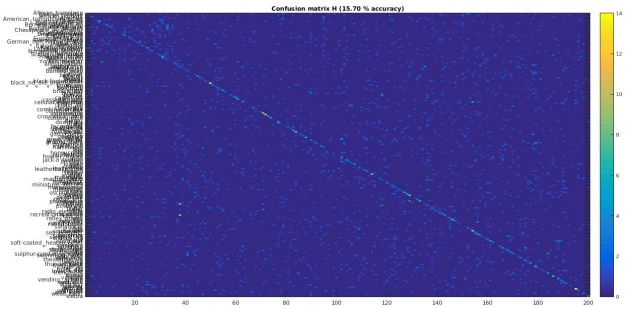

Figure 7. Confusion Matrix changing the X and Y spatial configuration to [48]

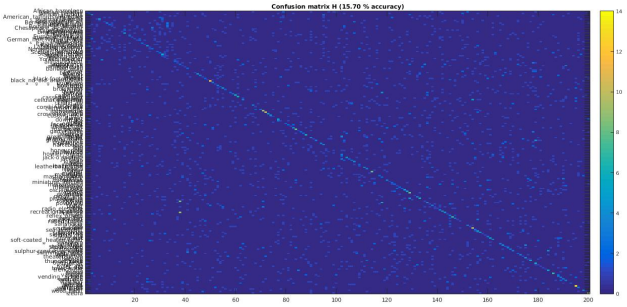

Figure 8. Confusion Matrix of ImageNet with default configurations

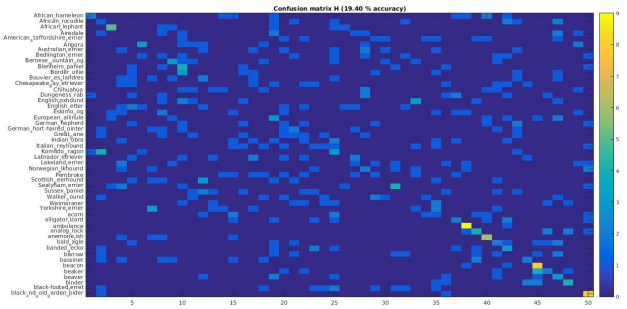

Figure 9. Confusion Matrix of ImageNet with default configurations

4. Discussion

PHOW uses the feature distribution as well as the spacial information of the features found in the image to generate an adequate representation. After obtaining the main features by means of SIFT and the k-visual words, the pyramidal spacial histogram of words adds the spacial components of the representation. In this way, PHOW offers a coherent image representation respecting type of features and location in the images such that, images with small histogram distance are similarly distributed respecting number and location of objects and features.

This characteristic allows a correct classification if the images from the same class are very similar, however, two images possessing the same number of representative features from a class with totally different spacial distribution of these features might be classified incorrectly, as their pyramidal spacial histograms will be alike for lower levels and very different in higher levels.

This is the case for Caltech101 and ImageNet. Caltech101 images from every class are very similar respecting orientation, clutter and pose[1]. Having one centered object per image with little clutter in a stereotypical pose assures a homogeneous spacial distribution of features and, consequently, a correct classification using PHOW and a SVM. However, this is not the case for ImageNet as the images of this dataset have a bigger variance in the object and features spacial distribution in the images. Having for the same class multiple and different types of the objects as well as different scenes containing the objects very distant histograms of visual words are expected, leading to a low classification performance by the SVM. This explains an ACA of 68.1% for Caltech101 and 15.7% for Imagenet.

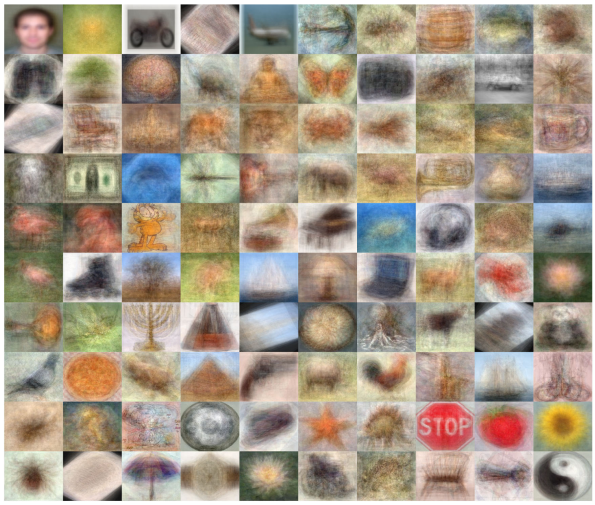

Figure 10. Antonio Torralba composed image averaging the images of each category of the Caltech101 database. Image taken from www.vision.caltech.edu/image_Datasets/Caltech101/#Discussion

Other characteristics of the image as scale, rotations or affine transformations of the features do not influence the representation and classification performance as one of the main advantages of the SIFT descriptor is that it is invariant respecting the above mention transforms. Scale invariance is given by the feature extraction in different resolutions using a difference of Gaussian (DoG) which allows the identification of representative features independently of the size it appears in the image[3]. Also, the rotation invariance is achieved by means of the gradient orientation in a local patch, which, for the same feature should be consistent between multiple instances as the spacial location of the patches and therefore gradient histogram is preserved in affine transformations.

SIFT itself corresponds to a descriptor of the local features based on the gradient histogram of local sub-patches. However, it does not incorporate the spacial location of the features in the image but the spacial location of the histogram of gradients in the feature. PHOW uses SIFT to incorporate the spacial location of the features in the images by concatenating the histograms of visual words at different resolutions. As PHOW uses SIFT to find and represent the images based on its' features at different resolutions using a spacial pyramid it can be said it is scale invariant.

The most important parameters of the PHOW strategy corresponds to the number of words used to build the visual vocabulary and the number of windows used to construct the spacial pyramid. The number of words should be sufficient enough to have one word per feature but not too big to have a considerable redundancy in the representation of the same features, leading to confusion. Respecting the number of windows, it should be considered that a small number would lead to bigger regions, a smaller descriptor and a small importance of the feature spacial arrangement while an increased number implies a bigger descriptor in which the feature spacial disposition has much more importance. The C parameters of the SVM shows no difference in performance.

The best parameters for Caltech101 are found where the number of train images and visual words are increased and the number of windows to construct the pyramid histograms are maintained small. It can be seen from 1 that the number of train images influence the most the actual performance of the algorithm while the number of words and the number of test images influence the less, as expected.

The number of words used in the experiments over Caltech101 are still sufficient for a correct classification, however, this parameter is expected to have an intermediate optimal value respecting performance. The number of windows also affects the ACA however, it is not affecting considerable the performance due to the characteristics of the dataset. Finally, the best performance is found when using 15 train images and 10 test images with 600 visual words and 4 windows per spacial resolution. However, it is expected that the best performance should be achieved by training with the whole train set and a reduced number of test images with 4 windows per spacial scale.

Respecting ImageNet, the best parameters found using 15 train images and 15 test images with a reduced number of classes. However, it is expected that the best parameters corresponds to using the complete train set, a reduced test set and the minimum number of spacial resolution and windows in the spacial pyramid. For future work, it might be interesting to evaluate the performance of the methodology

using the visual words histogram without incorporating the spacial pyramid as it could improve the performance over ImageNet.

The methodology does not present a particular confusion for certain classes in Caltech101. For all experiments the confusion matrices shows a random and homogeneous distribution of the errors. For ImageNet, the methodology presents a particularly good performance for *Beacon* and *Ambulance* classes which can be explained by the distinctive features of this objects like sirens and lights combined with a rectangular car in ambulances and a single linear a thin element with lights in beacons.

Different approaches can help to improve the classification in ImageNet. The first one involves changing the type of algorithm used for the classification. This means, that the method that differentiates the images change, at it can be more effective. Some other type of algorithms that can be used are k-means, decision trees, among others. On the other hand, the other approach that we can do is to generate a pre-processing of the images.

This means, that maybe in the way that caltech101 are using the data for training the model can be improved. If other characteristics of the image are taken, major information can be retrieved and that will help the classification algorithm to determine the class in a more effective way. Finally, a hierarchical algorithm can be used such that some similar classes like agglomerate all the different breeds of dogs in just one class labeled as dogs.

5. Conclusions

The SIFT descriptor can be used to obtain the relevant features of the images and describe them using the concatenation of the histogram of gradients of the patches. After that, a visual vocabulary by clustering can be constructed and the quantization of the features can be performed using a kdtree. Then, the spacial information of the features can be obtained by means of a spacial pyramid.

The representation of the images using the pyramidal histogram of visual words allows to represent the images based on its content and spacial distribution of features. This methodology gives a considerable weight to the spacial distribution of features in the images such that, images containing the same features in different locations will posses different PHOW representation and maximum histogram distances despite the same content. This explains the performance differences between the Caltech101 and ImageNet datasets.

As expected, an increased train size and visual vocabulary improves the performance of the classification. One of the most important parameters corresponds to the number of windows used to construct the spacial pyramid which can be seen, when reduced, improves the ACA.

One of the most important characteristics of the SIFT descriptor and the PHOW representation is that it is invariant to scale and affine transformations like rotations. This is achieved because the concatenated gradient histograms of features remain similar for different instances. However, one of the disadvantages corresponds to the increased weight given to the spacial distribution and the computational resources needed for a large descriptor.

The best ACA obtained using PHOW over the Caltech101 is about 68% and 19% over ImageNet. Some possible ways to improve the performance might be changing the classifier, weighting the visual words histograms such that low spacial levels posses higher importance respecting the higher levels or training with a reorganization of the images (by partitioning them into n parts and mixing them to form a new image) to reduce the spacial distribution predominance over frequency. Not all the classes will be distinguished but at least the *root class* will have a high recall.

References

[1] L. Fei Fei and P. Perona. Caltech101, 2006.

[2] S. A. Hussein, H. E. Naby, and A. A. A. Youssif. Image Multi-Classification using PHOW Features. *IOSR Journal of Computer Engineering*, 18(5):2278–661.

[3] F. Lu, X. Yang, R. Zhang, and S. Yu. Image classification based on pyramid histogram of topics. In *2009 IEEE International Conference on Multimedia and Expo*, pages 398–401. IEEE, 6 2009.

[4] d. Scikit. RBF SVM parameters scikit-learn 0.19.1 documentation, 2017.

[5] Stanford Vision lab. ImageNet, 2016.

6. Code and Images

The code is available at the team's repository at *https://github.com/steff456/IBIO4680/tree/master/08-PHOW/Answers*