Pyramid Histograms of Visual Words (PHOW)

Natalia Andrea Durán Castro
Universidad de los Andes
na.duran@uniandes.edu.co

Ana M. Velosa Orduz
Universidad de los Andes
am.velosa@uniandes.edu.co

Abstract

This paper presents a comparison of PHOW over two datasets, Imagenet and Caltech 101. The method developed was based on the VLFeat method of PHOW used on Caltech101. The results shows a better performance to Caltech101, with an ACA of 61.05 percent.

1. Introduction

1.1. Datasets

Imagenet is a dataset of images that are organized according to a hierarchy presented on WordNet. This is a large database of lexical in English. It is grouped depending of conceptual-semantic and lexical relations [4].The Imagenet dataset categories consists of nouns in the synonym set as per WordNet. There are more than 80,000 synonym sets of nouns in WordNet [4]. For each category, the ImageNet dataset provides 10,000 images with human annotations and quality control [3].In the figure 1 is shown three images of this database.

Caltech 101 is a dataset of 101 different categories. In each category, there are between 40 to 800 images. The size of each image is around 300x200 pixels. Also, this dataset included anotations. According to the error reports of Caltech, the performance is better when all the images of test are used and, it is not normalized to all the categories [2]. In the figure 2 is shown images of this database.

1.2. PHOW

Pyramid Histogram of Visual Words (PHOW) is an unsupervised trained model based on the engineered models to use in computer vision [1]. This is an extension of the bag of words model (BoW), where the images are considered as words. In this way, a BoW model considers the images features as words. Then, the bag words would be a sparse vector(vector where the majority of values are zero) with these words [1]. Thereby,it is possible to known the frequency in which a word(image feature) is repeated on the image.

Figure 1. Images on the dataset ImageNet200.

Figure 2. Images on the dataset Caltech 101. From left to rigth, an image of the category airplanes,accordion and barrel.

However, the BoW method does not consider the spatial information. For this reason, the PHOW method subdivided the image into incrasiling sub-regions as pyramids. The histogram is computed in each one of these regions.

The scale invariant feature transform (SIFT) is an algorithm to detect and describe local features in images [1]. The difference between this algorithm and PHOW is that SIFT does not read the image features as words. Then, the PHOW can use the SIFT descriptors with k-means clustering to transform them in a visual vocabulary that can be used as words [1]. Due to SIFT is invariant to changes in scale and orientation of the images, the PHOW would be invariant to this.

The difference between textons and PHOW is that textons just considers the spatial information on a color space of representation. On the other side, PHOW takes account more information of the image features. This shows better results in large-scale datasets.

1.3. Methodology

To develop the PHOW, it was used the VLFeat Library developed by Andrea Vedaldi and Brian Fulkerson *wget http://www.vlfeat.org/download/vlfeat-0.9.21-bin.tar.gz* [5]. Then, it was decompress with the command *tar -xzvf vlfeat-0.9.21-bin.tar.gz*. On the other side, to acquire the ImageNet dataset was used the command *wget http://bcv001.uniandes.edu.co/imageNet200.tar*

The hyper parameters used in the method of this library are the number of images, number of classes, spatial information in reference to X or Y and the features of the image (words). The evaluation method in this case is SVM with a chi-square kernel [5]. To choose the best hyper parameters to each dataset is necessary to know the number of class in train an test. In the case of Caltech there are 101 categories, and there are 200 in ImageNet. On the other hand, the spatial information, number of words and the number of class to train and test are going to be variable. Also, it is possible to change the number of categories to test the algorithm in a tiny dataset.

2. Results and discussion

2.1. Caltech

In the figure 3 we can see the value of the ACA that was obtained with the Caltech database with all the images. On the other hand, in the figure 4 we can observe the value that the ACA takes with the imageNet200 database as if it were a small problem, in which a smaller amount of class is taken.

We can see that the accuracy of imageNet200 is much lower than that of Caltech which is expected because the algorithm was made to classify the Catech101 dataset. The parameters that were changed each time the algorithm was run were the number of images that were used for validation and for training, the value of the step and finally, what value that was given to the C parameter of the SVM. By modifying the value only of the step we could see that the value of the ACA did not change no matter how big or small we put the error range of C. On the other hand, when changes were made with respect to the step, we observed the same as with parameter C, which did not significantly change the ACA.

Finally, when the number of images in the validation and the interweaving was modified, it could be observed that

data set from Caltech.jpg

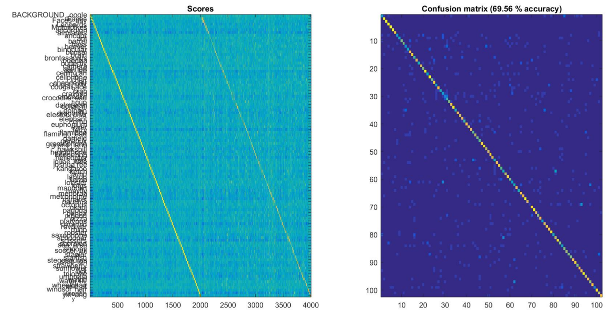


Figure 3. ACA of the complete Caltech Dataset without modifications

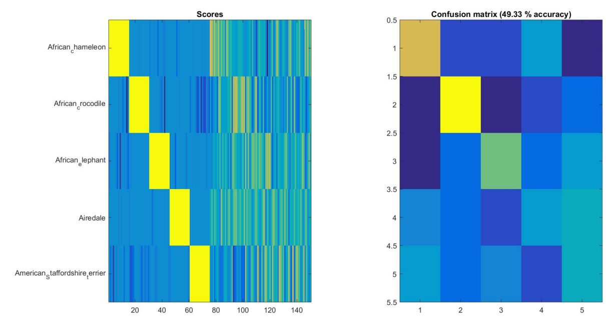
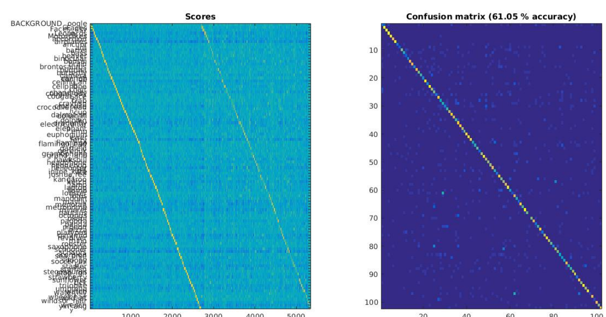


Figure 4. ACA of the tiny problem of the imageNet200 DataSet

test train a 30.jpg



e

Figure 5. ACA of the complete Caltech Dataset with resize of the amount of images in train and test to 30

for more images and for less images the value of the ACA diminished, as can be seen in 5 and6

Taking into account all of the above, it is important to emphasize that a change of 20 images in the validation and test images generates a very significant change since it goes from an ACA 61.05 to an ACA of 1.37 percent.

The num patialX / Y was also modified to dobre, that is, [4 8] but no change was observed in the ACA obtained as observed in the figure 7. On the other hand, in the figure 8 is shown the evaluation to the algorithm with 100 images size of train and test to 10.jpg

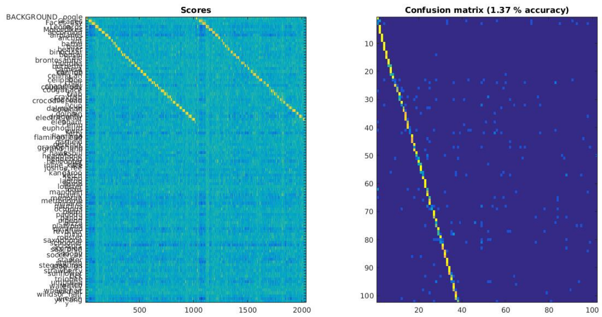


Figure 6. ACA of the complete Caltech Dataset with resize of the amount of images in train and test to 10

on the numSpace.jpg

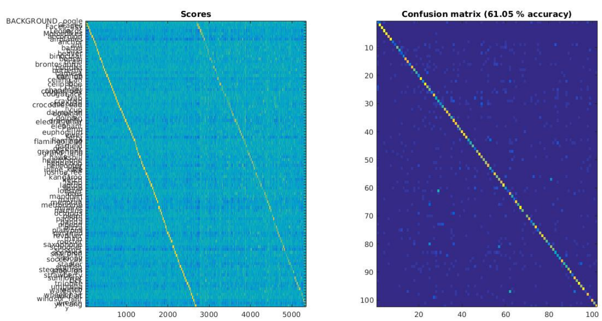


Figure 7. ACA of the complete Caltech Dataset with resize of the numSpatialX/Y to [4 8]

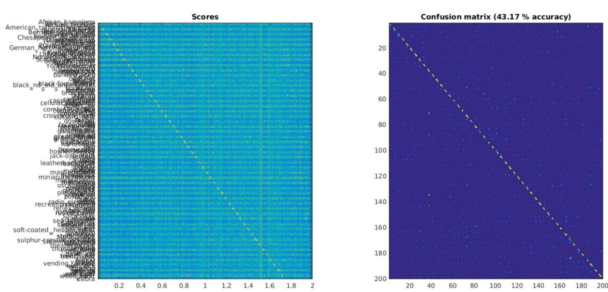


Figure 8. ACA of the complete Caltech Dataset with resize of the numSpatialX/Y to [4 8]

in train and 15 in test. This value of ACA in this case is 0.4317.

In conclusion, this algorithm can be use to classify other datasets if adjust the parameters. Principally, the parameters of number of images to train and test.

References

[1] S.-M. Khaligh-Razavi. Mrc cognition and brain sciences unit,cambridge university,cambridge, uk. pages –. 1, 2

[2] R. F. L. Fei-Fei and P. Perona. One-shot learning of object categories. IEEE Trans. Pattern Recognition and Machine Intelligence. In press. Retrieved from: *http://www.vision.caltech.edu/Image*$_Datasets/Caltech101/Description.$ 1

[3] O. R. J. K. J. D. A. B. Li Fei-Fei, Kai Li. Imagenet. Retrieved from: *image-net.org/about-people*. 1

[4] G. A. Miller. Wordnet. Retrieved from: *word-net.princeton.edu/people*. 1

[5] A. Vedaldi. Phow$_caltech101. Retrievedfrom : http://www.vlfeat.org/applications/caltech-101-code.html.$ 2