PHOW laboratory

Diego Martínez
Universidad de Los Andes
Cra 1N 18A-12 Bogota (Colombia)
da.martinez33@uniandes.edu.co

Felipe Torres
Universidad de Los Andes
Cra 1N 18A-12 Bogota (Colombia)
f.torres11@uniandes.edu.co

Abstract

Research in Computer Vision over the years has tried to find a propper way to represent Images according to different descriptors, through research many descriptors have been purposed such as HOG(Histogram of Oriented Gradient)[1], still in this implementation a dense SIFT [5] is used to extract the features of the desired images, then with the implementation of bag of words [4] a descriptor for images is generated. This methodology is applied in eight different experiments on caltech-101[2] database and on a small sample of imagenet database[3]. Highest accuracy is obtained in caltech with lower number of image classes to train in; lowest scores are obtained on imagenet. This result can be explained as images in caltech are catalog images, to put it in more common words, an image in caltech has the object class centered on it without any other classes, while in imagenet an image can contain several classes.

1. Introduction

Understanding visual information is an easy task for humans thanks to the huge training that our brain has had over the millions of years that we have inhabited this planet, even if we do not think about it too much the centerpiece of our neural network is a fine tuned machine that relies on its many layers in the visual cortex to process images and extract data just by paying attention for a moment; on the other hand this task is a real challenge when it comes to computers and machines that are by far not as nearly trained to do so as our brains.

To begin with, an image can be understood by many ways; for us human it is really easy to find Visual Patterns and rely on them to classify images, yet machines cannot do so in a fast and automatic manner such as we do because for machines an image is just a set of numbers in an array. On the other hand Visual Patterns can be edges, shapes, colors and textures to mention a few of them.

However, as patterns can be found in images, if an image is to be divided into its different locations and then each location is to be described according to its features; this method is known as SIFT. This method was first deviced by David Lowe in 1994, and it is invariant to scaling, traslation and rotation; partially to illumination and changes affine or 3D projection.[5] Image classification with the implementation of SIFT is done by extracting the main features of the image class and then creating a descriptor based on it; thus this descriptor can be described as a dictionary. Most articles and many texts that can be found along the many libraries and written information, images are composed of data that alongside a category is common. This kind of representation was defined by Zellig Harris back in 1954 , this technology is not restricted only to images; it began with text analysis and description [4]. To put an example, nowadays email technology can detect spam by reading the contents of a mail; this works because those mails contain about the same words everytime, say: Free, Gift, Viagra, Claim, Stock. These kind of words that are common for this kind of advertisement can be stored in a group and that would be a bag of words.

In Computer Vision a bag of words for a group/class is the information about shapes, textures that can describe said class; to put an example, the bag of words for a human face can be put in terms of the shape of the mouth, the nose, the eyes and cheeks. Should an image contain abut the same words as 'Cheek','Eyes', 'Mouth' then it would be an image of a face.

2. Methodology

Methods taken into account to employ PHOW as a representation feature for classification is decribed by several steps:

2.1. Database

In this study were used two different databases in order to understand the influence of database selection in predicted responces of a classifier. Thus, the datasets used are explained next:

- **Caltech-101:** This database was made by Fei Fei Li

when doing her doctorate in California Institute of Technology, back in 2005. This database contains 101 categories [2],and each category then contains from 40 to 800 images. It is important to notice that these images have the same size (300*200 Pixels) and they are catalog images meaning that there is only one class contained in one image.

This database is no longer being used to develop algorithms due to scandals that took place in the past, and also because as the images were really easy to classify, the problem is already solved in there with the state of the art algorithms.

- ImageNet:Is a database based on WorkNet Herierarcy of Princton University, but instead of groups of words, it has images. Unlike caltech 101, this database is formed by thousands of categories, with an average of five hundred images per category [3]. Moreover, images from this dataset are represented by a more diverse type of images that represent a category, which is totally different form catalog images, taken from usual internet images that people usually upload. For this study a small part of this data base were used, consisting of not more that 30 categories and 30 images per category.

As experiments to test the function of PHOW descriptors and classification made for Caltech 101, different sizes of database were implemented for both dataset.

- The number of categories used changed from 10 categories to 30 categories, in order to understand the effect of rising the number of categories in the performance of the algorithm
- The number of images per category was also changed. As an other variable of the experiment we take 10 and 30 images per category, which was useful to understand the effect of the database's size in the performance.

These was done for Train and Test datasets of main databases.

2.2. Image Representation

Image representation for both of the databases was done with the implementation of Bag of Words for each image; the descriptors extraction was done with the implementation of dense Sift descriptors.The sift implemented is the one found in Andrea Vedaldi's library Vlfeat. To do so, we started from the script *phow_caltech101.m* written by Vedaldi. For the implementation in caltech 101 some lines were changed and as it is a function, input parameters were altered so that experimentation could be done with a single script in which the inputs could be varied automatically.

For Imagenet again the script developed by Vedaldi was implemented, only that this time more lines were altered as we have different directories and much more images.

2.3. Classification

Classification of images was depicted using a non linear model build by Support Vector Machines, which metrics is determined by Chi square distant. The library used to employ SVM was the VLFeat PEGASOS SVM solver from VLfeat open source library.

2.4. Evaluation methodology

After trainning stage, Classificator models were tested for each experiment, and then Confision matrix and Average Classification Accuracy were obtained.

Summarazing evaluation methodology, to test de Bag of words representation for both databases 8 different experiments were devised, several parameters could be changed and the ones that where altered where:

- The number of classes to implement the code in.
- The number of images to be used to develop the descriptor and test it in Train and Test datasets.
- the number of spatial partitioning of X and Y axis to obtain SIFT descriptors.

Note that one important parameter, the number of words per bag remains the same, being by default 300. To explain the tests we will begin making some contractions:

- Image Classes = ImC.
- Number of images used = Num Im
- SpatialX and Spatial Y = SX, SY.

Now that the conventions have been made, we ennumerate our eight different tests.

- **First Test** First Test was done using ImC = 10, NumIm = 10, SX = 2 and SY = 2.
- **Second Test** Second Test was done using ImC = 10, NumIm = 10, SX = 8 and SY = 8.
- **Third Test** Third Test was done using ImC = 10, NumIm = 30, SX = 2 and SY = 2.
- **Fourth Test** Fourth Test was done using ImC = 10, NumIm = 30, SX = 8 and SY = 8.
- **Fifth Test** Fifth Test was done using ImC = 30, NumIm = 10, SX = 2 and SY = 2.
- **Sixth Test** Sixth Test was done using ImC = 30, NumIm = 10, SX = 8 and SY = 8.

- **Seventh Test** Seventh Test was done using ImC = 30, NumIm = 30, SX = 2 and SY = 2.

- **Eigth Test** Eigth Test was done using ImC = 30, NumIm = 30, SX = 8 and SY = 8.

3. Results

For both datasets a confussion matrix was developed to check the performance of the methodology.

3.1. Caltech 101 database

Resultant confusion matrices fo every experiment is shown in figures 1 to 8.

3.2. ImageNet

In order to compare the effect of changing the database and analyze the performance of PHOW based classificator function made for Caltech 101, validation results for training stage were obtained too. Results of validation are shown in figures 9 to 16. Then results of evaluation in Test database are shown in figures 17 to 24.

4. Discussion

Because were used two different databases to test the performace of same algorithm, firstly, they must be analyzed separatedly.

In the case of Caltech 101, classification accuracy had a high value with a maximum of 83% and a minimum of 58%. This variation between experiments describes the effects of changing number of catergories, size of training and testing dataset and spatial partitioning. Thus, a constant pattern is shown by rising the level of each variable which is a decrease in accuracy. In other words, increasing changing number of catergories and size of training and testing datase makes the accuracy lower. This could be the result of the raise in variation in descriptors and instances, that are the images. This is that a large amount of images increase variation of instances which obstruct clean and correct building a hyperplane for SVM model classification.

In the case of spatial partitioning, if increased, more exactly spatial information is added to the descriptor that is used to classify each image, and then accuracy is increased.

For ImageNet results were lower with a maximum accuracy of 39% for validation and 13% for testing, which is a direct result of variability in images of each category from this database. In this case spatial partitionig variation doesn't have a appreciable effect. Also, a greater amount of categories reduced the accuracy. However, worth to mention that increasing the number of images per category makes the classifier more accurate. This could have sense in that if this database have more varied images, increasing the number of images also increase the caracteristics to be acquire by the model made by SVM, making it more robust to variation in images.

5. Conlusions

Image recognition based solely on descriptors is a hard task when it comes to natural images, by natural images we reffer to those that can be found on nature and not on catalogs; the task gets complicated as we can have more than two classes within the image, therefore extracting the descriptors for said image class would take into account the other classes too.

SIFT as a descriptor extraction works well for caltech 101 database, however it starts to get limited should we increase the number of classes and/or number of images to be taken into account when generating the descriptor; the computer can just simply run out of memory.

Wishing to get better results of ACA in imagenet, the descriptor could be done while taking into account extra information as textures and spatial information; however it is important to know that a method done in this way could very easily finish off the memory of the computer while running on a database, also bear in mind that it would be extremely slow.

References

[1] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*.

[2] L. Fei-Fei, R. Fergus, and P. Perona. ne-shot learning of object categories. *IEEE Trans. Pattern Recognition and Machine Intelligence. In press*.

[3] L. Fei Fei, L. Kai, O. Russakovsky, J. Krause, J. Deng, and A. Berg. ImageNet.

[4] Z. S. Harris. Distributional structure. *Papers on Syntax*, pages 3–22, 1981.

[5] D. Lowe. Object recognition from local scale-invariant features. *Proceedings of the Seventh IEEE International Conference on Computer Vision*, 1999.

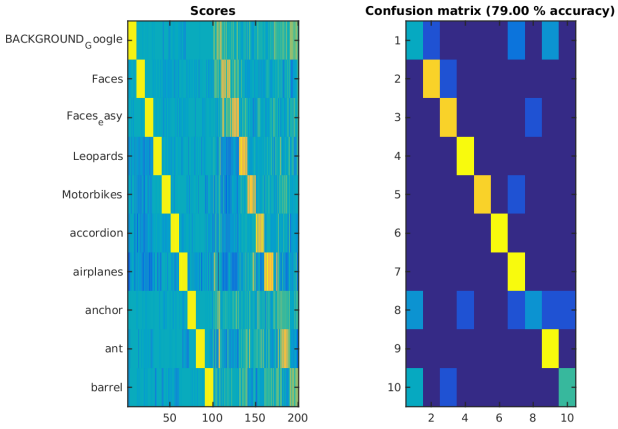

Figure 1. Confusion Matrix for 10 categories, 10 images per category and abith axis spatial partitioning of 2, in caltech 101 database

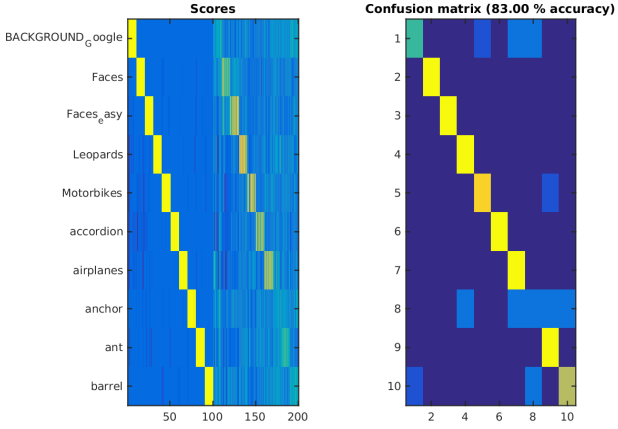

Figure 2. Confusion Matrix for 10 categories, 10 images per category and abith axis spatial partitioning of 8, in caltech 101 database

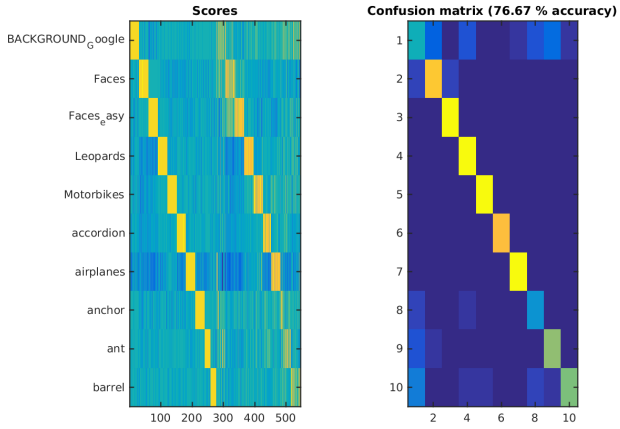

Figure 3. Confusion Matrix for 10 categories, 30 images per category and abith axis spatial partitioning of 2, in caltech 101 database

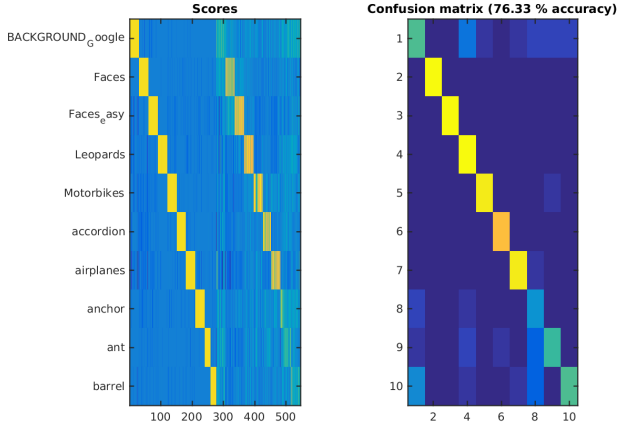

Figure 4. Confusion Matrix for 10 categories, 30 images per category and abith axis spatial partitioning of 8, in caltech 101 database

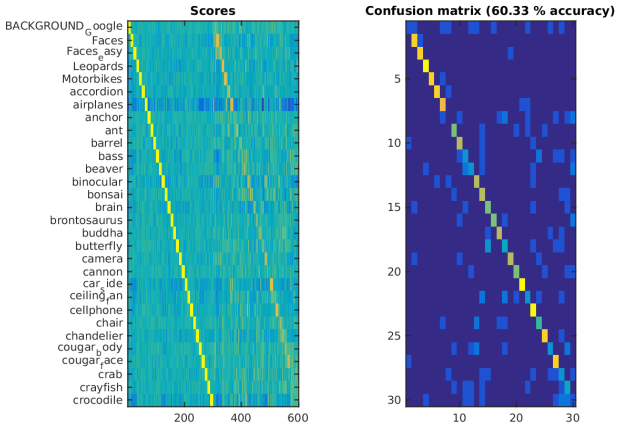

Figure 5. Confusion Matrix for 30 categories, 10 images per category and abith axis spatial partitioning of 2, in caltech 101 database

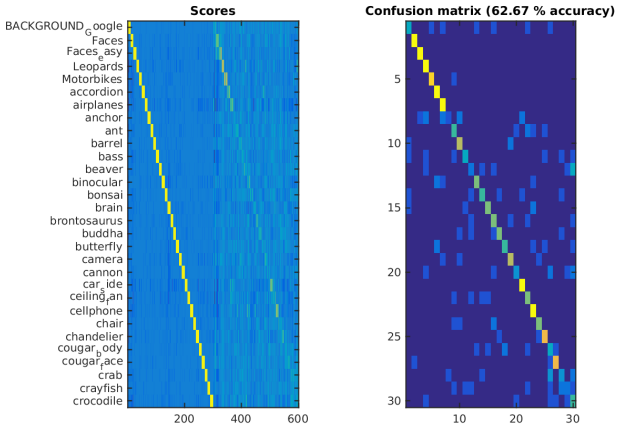

Figure 6. Confusion Matrix for 30 categories, 10 images per category and abith axis spatial partitioning of 8, in caltech 101 database

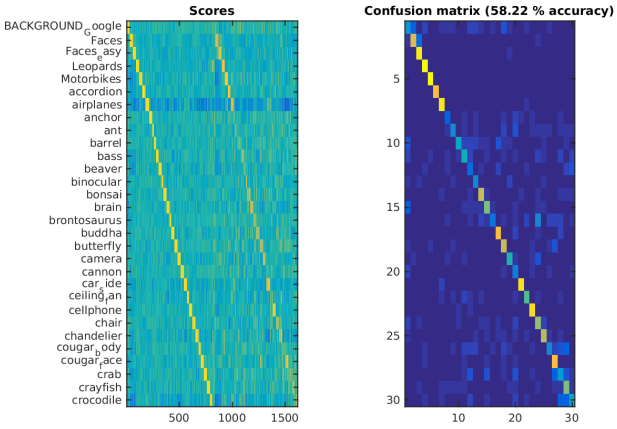

Figure 7. Confusion Matrix for 30 categories, 30 images per category and abith axis spatial partitioning of 2, in caltech 101 database

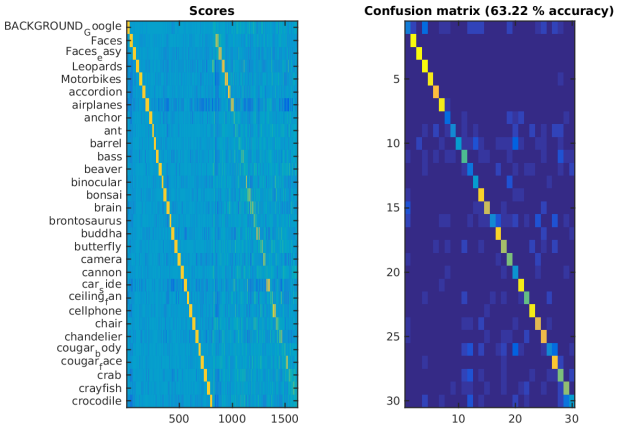

Figure 8. Confusion Matrix for 30 categories, 30 images per category and abith axis spatial partitioning of 8, in caltech 101 database

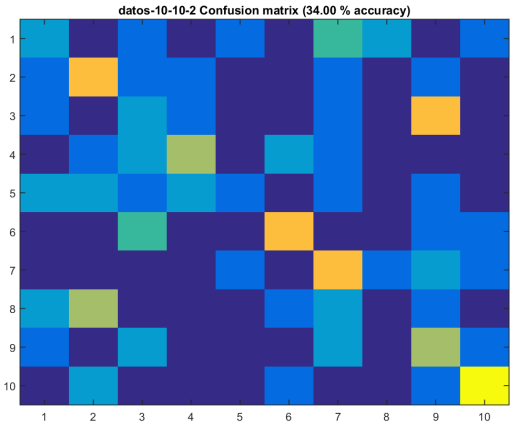

Figure 9. Confusion Matrix for 10 categories, 10 images per category and abith axis spatial partitioning of 2, in caltech 101 database

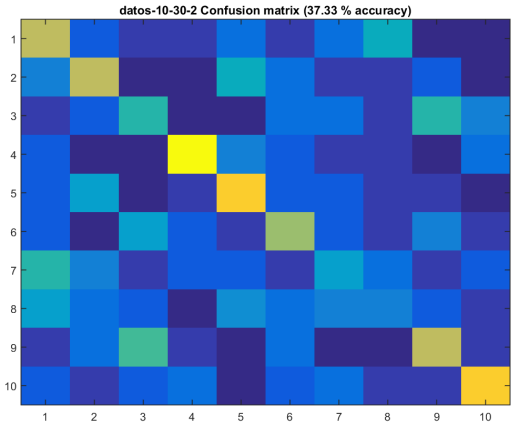

Figure 11. Confusion Matrix for 10 categories, 30 images per category and abith axis spatial partitioning of 2, in caltech 101 database

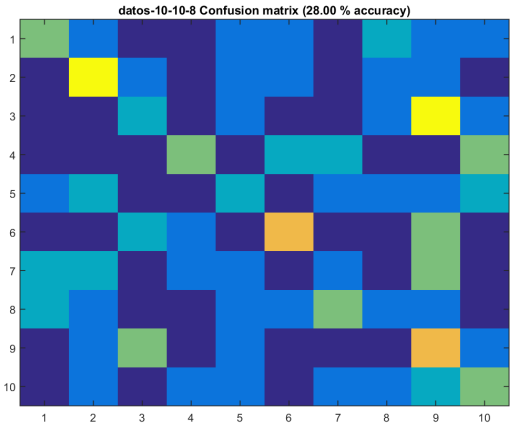

Figure 10. Confusion Matrix for 10 categories, 10 images per category and abith axis spatial partitioning of 8, in caltech 101 database

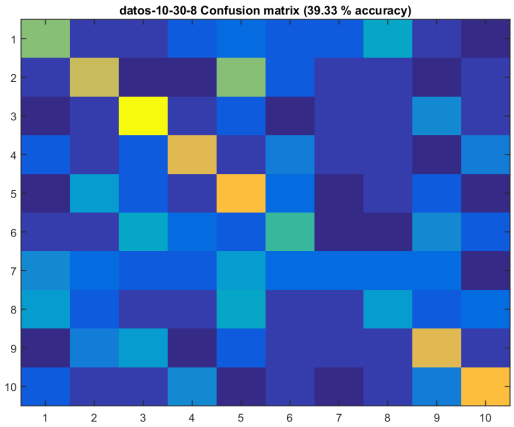

Figure 12. Confusion Matrix for 10 categories, 30 images per category and abith axis spatial partitioning of 8, in caltech 101 database

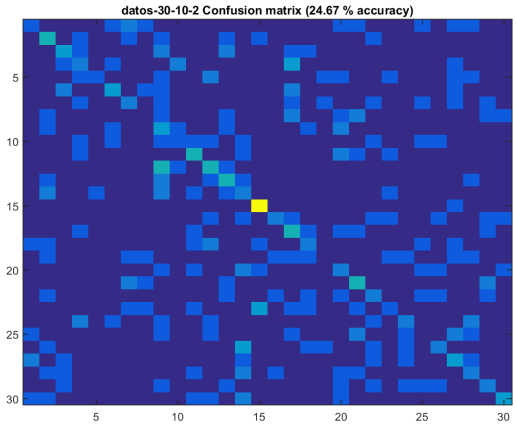

Figure 13. Confusion Matrix for 30 categories, 10 images per category and abith axis spatial partitioning of 2, in caltech 101 database

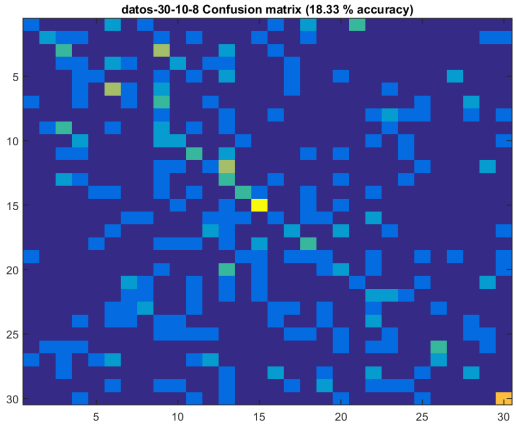

Figure 14. Confusion Matrix for 30 categories, 10 images per category and abith axis spatial partitioning of 8, in caltech 101 database

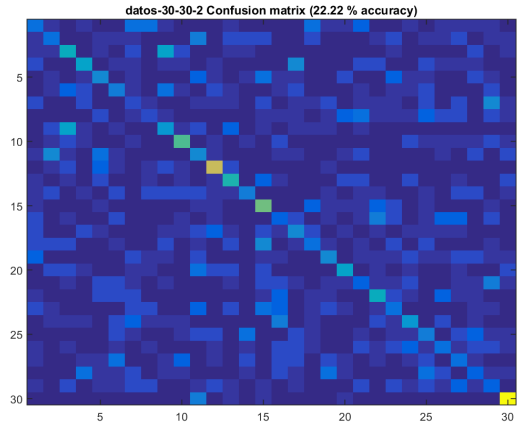

Figure 15. Confusion Matrix for 30 categories, 30 images per category and abith axis spatial partitioning of 2, in caltech 101 database

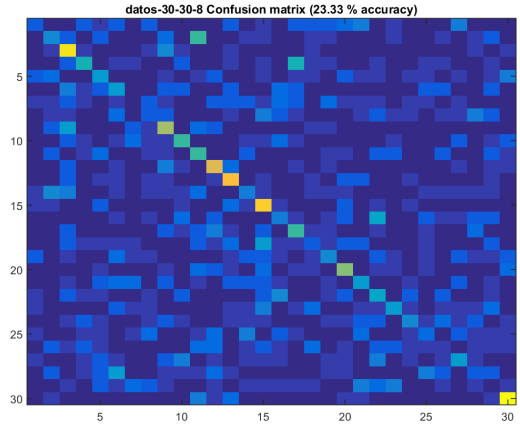

Figure 16. Confusion Matrix for 30 categories, 30 images per category and abith axis spatial partitioning of 8, in caltech 101 database

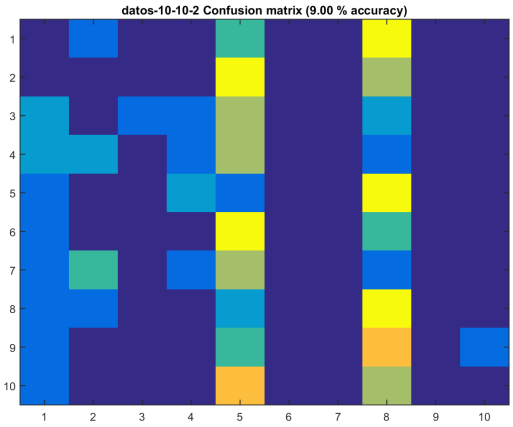

Figure 17. Confusion Matrix for 10 categories, 10 images per category and abith axis spatial partitioning of 2, in caltech 101 database

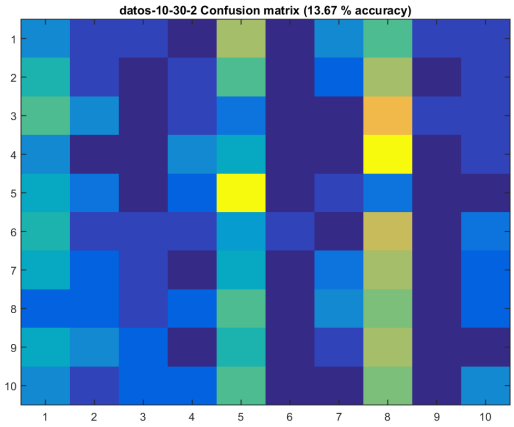

Figure 19. Confusion Matrix for 10 categories, 30 images per category and abith axis spatial partitioning of 2, in caltech 101 database

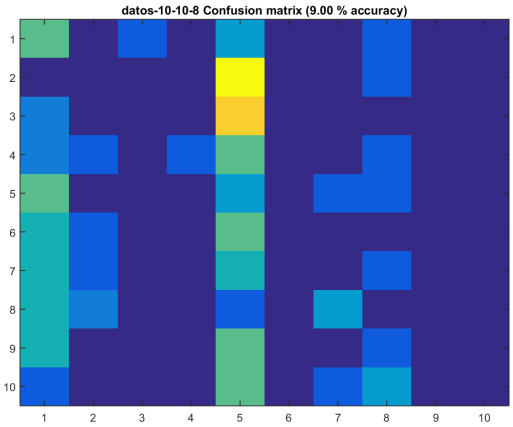

Figure 18. Confusion Matrix for 10 categories, 10 images per category and abith axis spatial partitioning of 8, in caltech 101 database

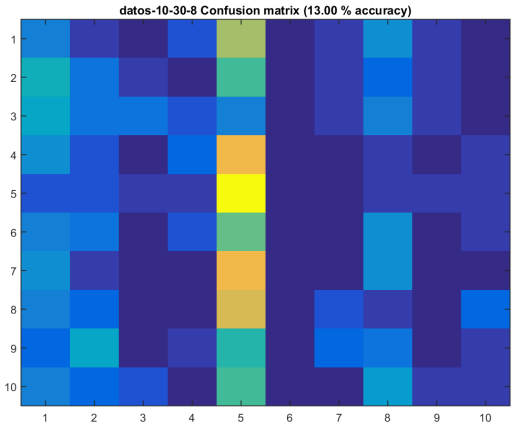

Figure 20. Confusion Matrix for 10 categories, 30 images per category and abith axis spatial partitioning of 8, in caltech 101 database

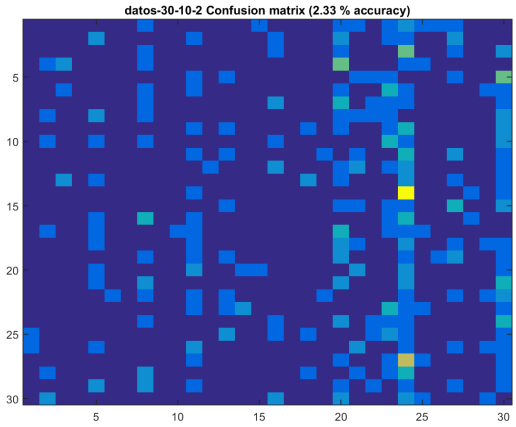

Figure 21. Confusion Matrix for 30 categories, 10 images per category and abith axis spatial partitioning of 2, in caltech 101 database

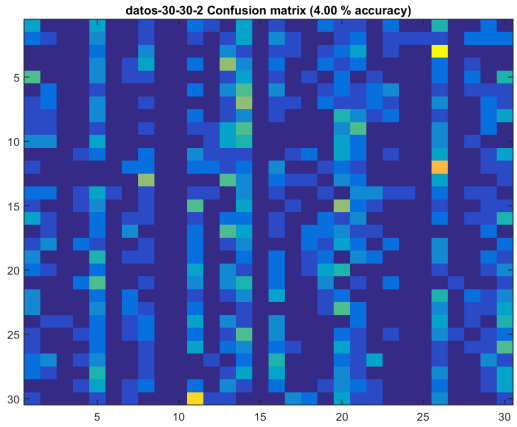

Figure 23. Confusion Matrix for 30 categories, 30 images per category and abith axis spatial partitioning of 2, in caltech 101 database

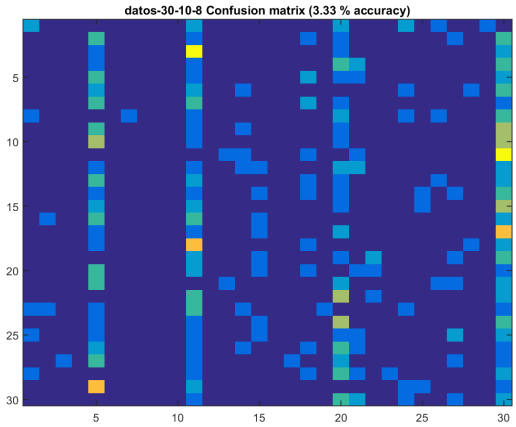

Figure 22. Confusion Matrix for 30 categories, 10 images per category and abith axis spatial partitioning of 8, in caltech 101 database

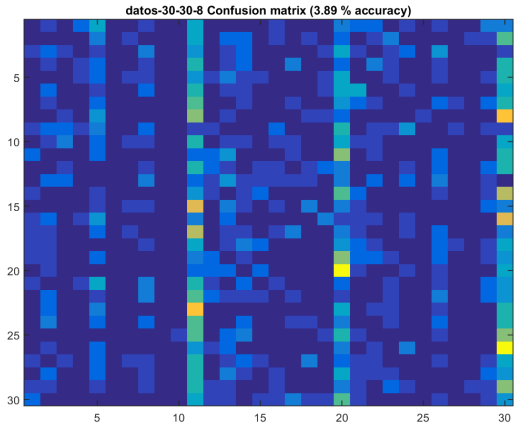

Figure 24. Confusion Matrix for 30 categories, 30 images per category and abith axis spatial partitioning of 8, in caltech 101 database