Pyramid histograms of visual words (PHOW) for Image Classification

Mariajosé Serna Ayala
Universidad de los Andes
m.serna10@uniandes.edu.co

Henry Daniel Torres Acuña
Universidad de los Andes
hd.torres11@uniandes.edu.co

Abstract

Image classification on computer vision problems have three basics phases: Training, validation and testing. Inside these stages there is always a process to follow: representation of the training images, the construction of the predictive model for the new images, the representation of the new images (test images) and finally their respective classification into one of the categories defined by the model. In the series of experiments presented here, we test the method of Pyramid histograms of visual words(PHOW) for image classification in the dataset Caltech 101, and a subset of the dataset ImageNet. For this purpose we used the VL_Feat library created by Andrea Vedaldi, and an example script included in the library page. The representation was made with PHOW which is based on dense SIFTs and the construction of a dictionary using this representation space. And the classification is made training an SVM. The best result, running over all categories for Caltech 101 was an ACA of 0.7. And for Imagenet an ACA of 0.006.

1.. Introduction

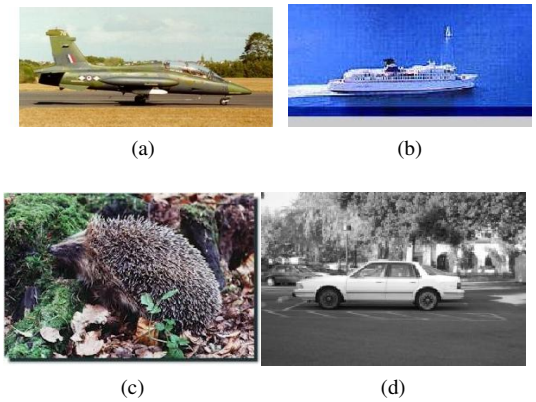
Image classification purpose is to set a label to an input image based on its contents. This is done by representing a set of train images with some attributes that can be calculated with some of the image information. Later, a classifier is trained, which means that with a set of images and its respective labels, it acquires the capacity to classify a new image with a certain probability (which depends in the training of the classifier, and the quality of the representation). As it can be predicted, the training of classifiers will depend on the number of images that are used to train and to the classifier that is used, where some have proven to be more efficient to others. Even though, increasing the number of images is going to also increase the time the computational time. Therefore, the improvements are in the representation phase.

The classification used during the set of experiments presented here is based on Pyramid histograms of visual words. This algorithm consists on two phases, as most common classification algorithms: training and testing. The representation is done by the Scale Invariant Feature Transform (SIFT), then by applying K-means to the images represented in the SIFT space a visual words dictionary is constructed. Where the center of every cluster k_i represents a visual word. Afterwards, for every image it is calculated an histogram with the frequency of every visual word. These histograms are the final representation of the image and these are passed, with the respective labels to train a Support Vector Machine (SVM). These method was proved in two important classification datasets Caltech 101 and imageNet.

Caltech 101 is a dataset of digital images, specialized for image classification, created in September 2003 at the California Institute of Technology. It has 101 categories, 102 including the background as a category. All categories have more or less 50 images, and a constant image size of 300×200. Every image is meant to have little or no background clutter. Also, the objects tend to be centered in each image, and most of them are presented in a stereotypical pose[2].

ImageNet is a much more recent Dataset. It was released in 2012 by the same person leading Caltech 101, Fei Fei Lee. This dataset, unlike Caltech, is divided into train and test images. We worked with a subset with 200 categories, although, the whole dataset has more than $1,000$ images. ImageNet is organized according to the WordNet hierarchy. WordNet is a large lexical database of English. 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, where the majority of them are nouns. In ImageNet, the purpose is to provide on average 1000 images to illustrate each synset[3]. Images are annotated for classification and recognition. For classification the images are divided into

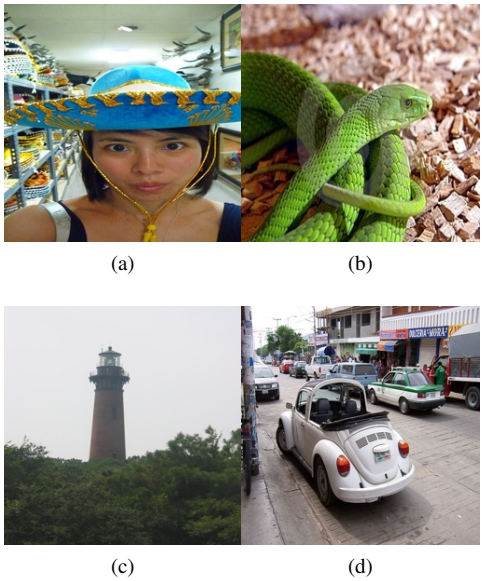
folders with the names of the different classes. In the subset we used there was 100 hundred images per class.



(a) (b)

(c) (d)

Figure 1: Example of Caltech 101 dataset images for classes: (a) Airplanes, (b) Ferries, (c) Hedgehog and (d) Cars. It can be seen that in all of them the objects are centered in the image.



(a) (b)

(c) (d)

Figure 2: Example images of ImageNet dataset images for classes: (a) Sombrero, (b) Green Mamba, (c) Convertible, (d) Beacon. Here the first image could also be part of the class faces, even though we only generate one label per image.

2.. Materials and Methods

For training and testing the feature model we used the open source library *VLFeat* [4] which includes most of the state-of-the-art algorithms for instance recognition such as Fisher Vectors, VLAD, SIFT, SVM classifier, k-means, etc. The library is based on C scripts for efficiency and compatibility and it provides a MATLAB interface ready to use. As an approach for the recognition problem we used the Piramidal Histograms of Visual Words (PHOW) algorithm contained in the library (specifically on the $phow_caltech101()$ function).

The PHOW strategy we followed is based on the combination of the construction of a visual vocabulary and a SVM classifier method. In this we encountered six main parameters that influence the performance of the algorithm. First, the number of images used for training and test per category limit the problem and allow us to make a sort if subsampling considering the size of the dataset. Then, the number of classes and the number of visual words are important to the accuracy of the algorithm either by adding precision to the categories the SVM can classify into or building a richer representation space, respectively. Finally, the x and y spatial partition and the C parameter of the SVM. Worth mentioning that the C parameter models confidence on the classification given by the distance to the formed hyperplane, thus tunning this parameter can make the SVM more relaxed on more restrictive. Another parameter we found it was important to the algorithm was the kernel of the SVM, since the SVM relies on the geometry of this kernel for dividing the data.

Arguably the core part of the PHOW algorithm is the construction of a quality representation space. For this, Vedaldi uses a dense SIFT to represent each image, this is a concatenation of the responses to multiple visual words at different scales in the partition of the images. These visual words are acquired by applying k-means[**?**] grouping algorithm to the descriptors obtained from the scaling. Then, it is necessary to build the spatial histograms with the number of hit for each visual word at all the used scales, this is the true descriptor of each image. Finally, a $CHI2$ SVM is used to construct the hyperplane that divides each category using the previously constructed histograms as support vectors. For testing the model it is necessary to construct the PHOW for the each new image and the used the previously trained SVM to classify the image. It is important to note that rather than a single SVM it is necessary to train at least as many SVMs as the given number of classes.

To run the classification in the Caltech 101 dataset we changed 6 different parameters: Number of words from the visual words dictionary, the number of images used for train, the number of images used for test, the SVM C, the number of categories taking into account and the spatial partition in X and Y. Based on these final parameters we

tried them for the ImageNet dataset.

3.. Results

Caltech 101

The first three experiments, changing the number of visual words (clusters), was run with the default values for the rest of parameters(5 categories, 15 train images, 15 test images). Tables 1, 5, and 3 show the changes in the parameters and the respective Average Classification Accuracy (ACA) from the confusion matrices obtained.

Number of Words	ACA
300	0.92
200	0.92
250	0.91
100	0.86
50	0.82
400	0.9
600	0.9

Table 1: Changing the number of visual words. The bold number corresponds to the original value.

SVM C	ACA
2	0.92
7	0.92
13	0.91
18	0.86
10	0.82
25	0.9
50	0.9

Table 2: Changing the value of the parameter c from the SVM clasiffier. The bold number corresponds to the original value.

Spatial Partition X	Spatial Partition Y	ACA
2	**2**	0.92
6	6	0.96
8	8	0.94

Table 3: Changing the value of Spatial partition x and y. The bold numbers correspond to the original values.

When changing the number of train and test images there was not a lot of options because not all categories had 50 images, some had less and therefore we needed to make sure that the sum of both train and test where not bigger than 40. These where tested with original values for the rest of parameters.

Number of test images	ACA
15	0.92
20	0.92
25	0.91
30	0.86
10	0.82

Table 4: Changing the number of test images for each category. The bold number corresponds to the original value.

Number of train images	ACA
15	0.91
20	0.906
25	0.95
30	0.96
10	0.90

Table 5: Changing the number of train images for each category for the training of the model. The bold number corresponds to the original value.

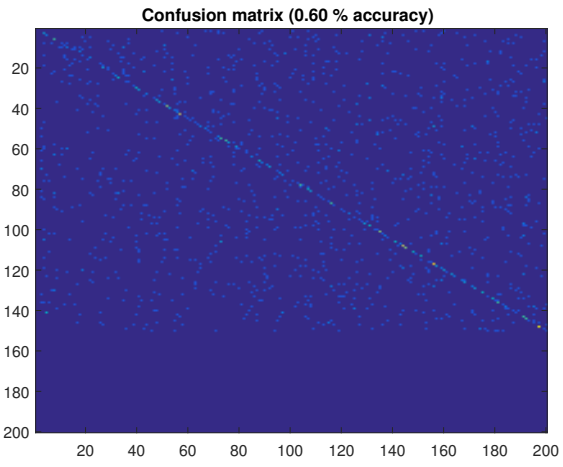
For the experiment changing the number of categories (Table 6) we used the number of train and test images that in the two previous experiments appear to be the optimal: $Trainimages = 30$ and $Testimages = 10$.

Number of categories	ACA
5	0.96
15	0.76
30	0.69
60	0.66
102	0.70

Table 6: Changing the number of categories evaluated. The bold number corresponds to the original value.

ImageNet

Training and testing the model for the ImageNet database based on the code and parameters used for Caltech 101 result in an ACA of 0.006 on the test set(Figure 3). Even though this only was calculated for 140 categories because every time we tried to run the trained model to classify test images the server would not let the process continue further than this.



(a)

Figure 3: Confusion matrix for whole set of test images from ImageNet dataset.

4.. Discusion

For Caltech 101 classification we suspected that by passing a larger quantity of images to train the model the classifier will be more robust. Even though we incurred in the error of not testing different parameters possibilities with a larger number of classes, or even the whole set. Because is evident that when taking into account a larger number of categories the probabilities of falling where going to be bigger.

But, the parameters that are more important for the PHOW strategy are the number of words and the value of C from the SVM classifier. In contempt of that the fact that we did not try different number of words with a higher number of categories, and therefore is not so relevant the changes obtain, it was a parameter that modifies the final ACA and it seems to reach to a point where the increment of the number of words would not increase the ACA but, on the contrary, it starts to decay. This also happens with the C parameter of the SVM. The C is the coefficient of the summatory of the errors allow on the margin between the support vectors for each category. Thus, the manipulation of the C parameter is a way to control the propagated error in the classification process. A high value of C leads toward a much restrictive classifier and a low value to a more relaxed one, influencing directly the accuracy of the classifier.

Also, other set of parameters present in the script, and that influence the performance of PHOW are the scales and a parameter called 'Step'. The scales could not be changed because these are defined inside the function vl_phow(), but the scales are the ones defining at which resolutions the image is going to be represented. And 'Step' consists in the size of the grid (in pixels) at which the dense SIFT features are extracted. These two parameters, as the number of words, are the ones with bigger influence in the final performance of the algorithm, because all of them influence on the image representation. Therefore, defining the correct space so it can be evident to the model which images are different and which ones are enough similar to belong to the same class.

In Caltech 101 there was not much of a difference between the classes, because as mention before, all their images have special properties like: centered objects in the image, regular backgrounds.. etc. But in ImageNet dataset the images have different categories inside of them, which can lead to errors. Also the objects are presented in different positions, which is also difficult, and even more whit dense SIFT representation, because its a lot based on the form of the objects. And when an object appear in a different perspective it forms changes a lot.

5.. Conclusions

- Although PHOW has proven to be a method with pretty good results on the experimental frame of *Caltech101* database we couldn't transfer effectively the knowledge on parameters to much complicated *ImageNet* database arguably because of the elevated computational effort that is necessary to complete the experiment with the optimal experimental parameters acquired on *caltech101*.
- The PHOW is a refined strategy for the recognition problem since it uses an information and dimension-rich representation space to describe the input images. However, the use of k-means grouping method to construct the visual words themselves incorporates an unnecessary and undesirable error from the very beginning of the algorithm. One way to solvent this flaw without going to an extreme over complication is to use a Gaussian Mixture Model for this purpose.
- Due to the 2008's computational constraints the implementation used a $CHI2$ SVM to reduce down the cost of training the classifiers. Anyhow, since the descriptors are basically histograms we propose to use a more appropriated intersection kernel instead.

References

[1] R. Szeliski. Computer Vision: Algorithms and Applications. Chapter 5:Segmentations. 2011

[2] L. Fei-Fei, R. Fergus and P. Perona. Learning generative visual models from few training examples: an incremental Bayesian approach tested on 101 object categories. IEEE. CVPR 2004, Worksh

[3] L. Fei-Fei and O. Russakovsky, Analysis of Large-Scale Visual Recognition, Bay Area Vision Meeting, October, 2013.

[4] A. Vedaldi and B. Fulkerson, "VLFeat : An Open and Portable Library of Computer Vision Algorithms",2008, http://www.vlfeat.org/