# Evaluation of PHOW method in Caltech 101 and ImageNet dataset

Saul Gómez
Universidad de Los Andes
sc.gomez11@uniandes.edu.co

Diego Valderrama
Universidad de Los Andes
df.valderrama@uniandes.edu.co

## Abstract

*Computer vision imaging is a complex task that can be tackled from different perspectives. The extraction of features using Scale-Invariant Feature Transform (SIFT) and its variants has allowed the development of methods that have improved the performance in the classification of multicategory images. Pyramid Histograms of Visual Words (PHOW) makes use of dense SIFT for the extraction of important features of images, which allows to create a dictionary of visual words that allow to classify a query image by its similarity with some of the words of the vocabulary. The current study evaluates the performances of PHOW algorithm in Caltech101 and ImageNet datasets. It was obtained an ACA of – and – for the train sets of Caltech101 and ImageNet, respectively. Finally, the performance of the method on ImageNet test set provides an ACA equals to –*

## 1. Introduction

Image classification is a complex process that may be affected by many factors [11]. It differs from traditional problem of image retrieval because images are classified under particular fields and classification systems are generally predefined [5]. For this reason, and joint with the high dimensionality of the feature space, traditional classification approaches generalize poorly on image classification tasks, which explains why methods train with one database do not show same performance on different datasets [2]. Several studies employed color histograms as an image representation because of the reasonable performance that can be obtained in spite of their extreme simplicity, performing generic object classification with a "winner takes all" approach, i.e. find the one category of object that is the most likely to be present in a given query image [2], [17].

However, classifying remotely sensed data into a thematic map remains a challenge because the complexity of the landscape in a study area, selected remotely sensed data, and image-processing and classification approaches, which may affect the success of a classification [11]. Additionally, such classification techniques based on low features (colors, textures, and boundaries) have been studied for years in the area of image retrieval and have been demonstrated that cannot successfully recognize objects at the same kind [7], [18]. In the same order of ideas, given the fact that an object could be recognized among other objects in a large image, the same object of multiple images could be recognized. Thus, keeping this into account, image classification could be addressed by considering various types of features to describe the image contents [7].

The main characteristic of data classification is dealing with plenty of class labels and a small number of samples [7]. Also, image local features and descriptors play a pivotal role in various computer vision applications, such as image registration and object recognition [13]. These aspects make feature extraction and selection a vital strategy to guarantee reliable and meaningful results for data classification alongside different advantages such fewer data storage and computation cost [7], [15], [6], [12]. To achieve this, the content-based image classification do some semantic type classification to the images according to their visual features [5]. Firstly, visual features are extracted to represent and describe the image accurately employing image processing and analysis techniques. Further, supervised classification algorithms are trained to implement classification models necessary to classify a query image [5]. Notwithstanding, this strategy also have limitations. One is evidenced when matching features across different images appearing in different scales and rotations, reason why Scale-Invariant Feature Transform (SIFT) has become one of the famous tools to dealing with it [10].

SIFT is a high probability object detection and identification method, which is done by matching the query image against a large database of local image features [10]. The main idea of SIFT is to carefully choose a subset, representative of the original, of the features and process it [10] Certainly, there are applications in which SIFT is not enough and working on a dense set of features, rather than the sparse subsets, represents a more efficient approach because provides more information than the corresponding descriptors evaluated only at selected interest points [10]. This process is often called Dense-SIFT. In the case of object cate-

gory or scene classification, experimental evaluations show that better classification results are often obtained by computing the so-called Dense SIFT descriptors as opposed to SIFT on a sparse set of interest points [1]. The dense sets may contain about 300 times more vectors than the sparse sets, which provides a better approach and a more accurate classification performance [10].

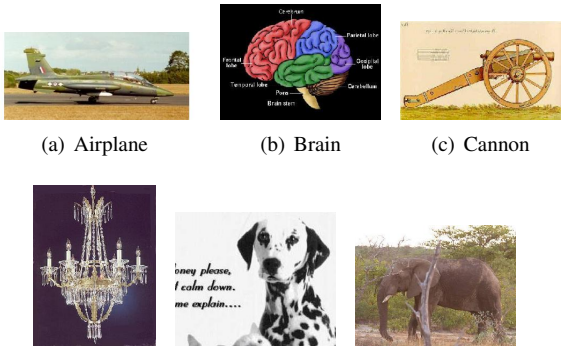
For this reason, the present study intend to evaluate the performance of Pyramid Histogram of Words (PHOW) strategy for image features extraction and further classification in two different database, Caltech101 and ImageNet. Supervised classifier was trained, the comparison between the results in both datasets was made and the relevant parameters of PHOW strategy were discussed. Finally, it was concluded regarding the limitations of the method used and the challenges that must be overcome in order to improve the results obtained.

## 2. Methods

### 2.1. Datasets

#### 2.1.1 Caltech 101

Caltech 101 dataset was created in 2003 by Fei-Fei Li, Marco Andreetto, and Marc Aurelio Ranzato at California Institute of Technology [4] . This dataset consist in 101 categories of natural images of roughly 300 x 200 pixels. Most categories have 50 images and the rest about 40 to 80 images [4]. It is important to clarify that Caltech 101 dataset is not divided in train and test set. Figure 1 shows some examples of the images included in the dataset.



(a) Airplane (b) Brain (c) Cannon

(d) Chandelier (e) Dalmatian (f) Elephant

Figure 1. Example of images for six image classes observed in caltech 101 dataset.

#### 2.1.2 ImageNet

ImageNet dataset was created by Stanford Vision Lab and is organized according to the WordNet hierarchy [8] . This dataset consist in 14.197.122 natural images divided in 27 classes. Each class has several number of subcategories with more than 400 images per each one and a total of images greater than 56K for class [8]. In this paper, we will use some images from 200 subcategories of different classes. Figure 2 shows some examples of the images included in the small subset of the ImageNet dataset used in this paper.

(a) African chameleon (b) Ambulance (c) Angora

(d) Beacon (e) Box turtle (f) Cannon

Figure 2. Example of images for six image classes observed in imageNet dataset.

### 2.2. PHOW method

Pyramid Histogram of Word (PHOW) originates from the Bag of Feature (BOF) and the combination of Space Pyramid Model. Bag of Feature is one of the popular visual descriptors used for visual data classification and is inspired by a concept called Bag of Words that is used in document classification [9]. In computer vision, a bag of visual words of features is a sparse vector of occurrence counts of a vocabulary of image features [9]. Now, feature detection and description is the first step of PHOW. To do this, a variant of the SIFT descriptor, proposed by Lowe, dense SIFT, captures invariant local image structural information. Then it creates a local descriptor for each key point detected by building an histogram of gradient orientations and image pixels in a window pacth centered at this point [9], [14].

Subsequently, local SIFT descriptors are extracted at regular grid points, rather than only at key-points [9]. When color images are processed, they are converted from RGB space to HSV color space with the SIFT feature extracted from each channel. For grayscale images, only the intensity is used. Hence, the resulting SIFT feature dimension is 128*3 for color images and 128 for grayscale images [7]. Once the SIFT features are obtained, “bag of words” model is used to quantize them into visual words by k-means clustering. Thus the image is represented by a histogram of visual word occurrences [7]. Every clustering center can be seen as a visual vocabulary and all vocabularies form a vocabulary dictionary [9].

### 2.3. Choosing the hyperparameters

Some experiments were made in order to obtain the hyperparameters with which we obtained the best performance using PHOW method in caltech 101 and ImageNet datasets. To do this, we change 7 hyperparameter, i.e, number of train images, number of classes, number of words, SVM conficende, steps, window size, NumSpatial. We began changing the number of train images and kept the rest of the parameters constant. We ran the algorithm to get the performance for each value of the range in which we change the hyperparameter and select the best one. After that, we kept the value found for this hyperparameter constant, changed the number of classes and the procedure was repeated for all of the hyperparameter mentioned. It is important to clarify that since ImageNet has a train set we use those images to obtain the hyperparameters. Table 1 shows the range between we changed all of the hyperparameters.

Table 1. Optimum hyperparameter values

| Hyperparameter | Range |
|---|---|
| Train images | 1-15 |
| Number of classes | 1-102 |
| Number of words | 300-800 |
| SVM confidence | 1-50 |
| Steps | 1-20 |
| Size | 1-20 |
| NumSpatial | 2-20 |

## 3. Results

We obtained results of the optimum hyperparameters for both datasets, Caltech 101 and ImageNet. We found that some of the hyperparameters are the same for the two databases and others are very similar. However, the greatest variation was obtained in the optimal number of words for each of the analyzed databases. The results of Table 2 shows that ImageNet dataset need a larger dictionary than Caltech 101 which means that was necesary used more clusters in ImageNet than in Caltech 101 dataset.

Table 2. Optimum hyperparameter values

| Hyperparameter | Caltech 101 | ImageNet |
|---|---|---|
| Train images | 17 | 19 |
| Number of classes | 10 | 10 |
| Number of words | 600 | 775 |
| SVM confidence | 10 | 10 |
| Steps | 5 | 3 |
| Size | 7 | 6 |
| NumSpatial | 2 | 2 |
| ACA | 0.9 | 0.467 |

After having the optimal hyperparameters, the number of classes was changed to 102 and 200 for Caltech and ImageNet, respectively. Then, we ran the algorithm in Caltech 101 to obtained the performance in whole dataset. For ImageNet, we first created a model using the train set and later evaluated the model in test set to get the performance of the method.

According to Figures 3, 7 the performance (ACA) of PHOW method is 62.2 % and 13.73% for Caltech 101 and train set of ImageNet dataset, respectively. The big difference in the accuracy could be because the differences in the features of the images in each dataset, i.e, Caltech 101 is a dataset in which the images have one center object assures a semi-homogeneus spatial distributions of the features for all images in a same class [4]. This feature allows the PHOW and SVM method to perform better in the Caltecth 101 data set than in a database where the objects are not centered and have different posese such as ImageNet dataset [8]. Table 3 sumarized the performance for caltech 101 dataset and for train and test set of ImageNet dataser.

Table 3. ACA performance of PHOW method

| Dataset | ACA |
|---|---|
| Caltech 101 | 62.2 % |
| Train set of ImageNet | 13.73% |
| Test set of ImageNet | 13.43 % |

## 4. Discussion

### 4.1. Relevant hyperparameters

As mentioned above, the number of words is an important parameter since it defines the size of the dictionary and is also related with the number of clusters used in K-means, making the classification dependent on this value. The confidence of the SVM is another important hyperparameter since it determines the importance of the margin that exists between the two support vectors allowing the performance of the algorithm to increase or decrease according to the analyzed data. The steps and the size of the sliding window are also important parameters because of from these values the algorithm obtained the pyramid histogram of word. On the other hand, how robust the algorithm is with specific dataset is related with the number of image train and classes making them important hyperparameters.

### 4.2. PHOW characteristics

As mentioned in methods section, PHOW strategy uses dense SIFT descriptor to extract visual features from the images. However, this method presents two major differences from the original algorithm developed by Lowe. First, the extraction of local features based on points is partitioned into two stages: keypoint detection and generation of feature descriptor [7]. Image description by Pyramid of His-

tograms of Visual Words (PHOW) method is an extension to the bag-of-words (BOW) model in which the extracted SIFT image features treated as words [7]. Yet, SIFT features extracted do not considers the local information feature of the image, thus, a spatial pyramid of visual words model is used to overcome the problem of dismissal of spatial information of local descriptors [3]. For this reason, the method considers both the global features (color and shape) and the local distribution information of the image, furthermore it is much more complete and flexible to describe the feature information of the image through multifeature fusion and the pyramid structure composed by image spatial multi-resolution decomposition, which improves the accuracy of image classification [5].

On the other hand, the SIFT algorithm, detects key points via a Difference of Gaussian (DoG) pyramid created using a Gaussian filtered copy of the image [7]. This means that SIFT keypoints are not influenced by a considerable lot of the complexities experienced in different techniques such as translation, rotation, scaling and also the noise effects [7]. However, the 128-measurements of the descriptor vector makes its feature extraction process relatively computationally costly [16]. In contrast, PHOW algorithm extracts features by using dense SIFT. In dense SIFT, there is no feature detection stage while local feature descriptors are extracted at every pixel to obtain a pixel-wise SIFT image. Specifically, for each pixel in an image, its neighborhood patch with a certain size like 48  48 is first divided into a 4  4 cell array. In each cell, an orientation histogram with 8 bins is used to quantize the gradient information. The histogram is constructed by accumulating the gradient orientations of all the pixels in a cell weighted by their gradient magnitudes. This denotes that the dense SIFT descriptor is not rotation and scale invariant since all the pixels in an image use a fixed-size patch as the neighborhood [13].

Additionally, unlike textons, where the response of each pixel to a filter bank was compared with a textbook product of algorithm training [19], in PHOW algorithm the image is partitioned level by level and each level of image is composed of several blocks [5]. Then a series of visual words histograms are formed for the representation of image from low resolution to high resolution in the feature space [5]. This has demostrated that guarantee reliable and meaningful results for classification task and computational efficiency [15], [6], [12], [7].

### 4.3. PHOW challenges

Based on the characteristics and limitations mentioned above the main challenge to overcome is the scale variant factor. Given that PHOW uses dense SIFT for image partitioning and include spatial information into consideration, it becomes very sensitive to translation, rotation or scaling transformations. Also, it should be noted that PHOW algorithm only takes into account a single feature representation, shape. This provides a bias on image features identification. Although PHOW system intend to include as much information of characteristic features of objects as can, intensity and texture information is being ignore, which can provide distintive data from objects that could improve classification accuracy.

### 4.4. Confusion classes

As can be noticed on figure 4, the algorithm performed less than 50% in approximately 33 out of the total 102, or approximately 33.66%. This indicates that in at least one third of the database, the algorithm is not able to correctly discriminate the images and assign them to their corresponding class. Also, of these 33 classes two were identified with a performance equal to 0, indicating that the algorithm is unable to distinguish these images among the other categories. One of these classes corresponds to octopus category, as shown in figure 4. Figure 5 shows different images that belong to octopus class on Caltech101. Here can be evidenced how different the images are between them, having different shapes from image to another due to the fact that some images are real octopus (see figure 5(b)) and other are just pictures or hand-drawn (figures 5(c), 5(d)). This particularity makes learning for the algorithm more complex and thus increases the error in classification. On the other hand, 13 out of 102 classes was perfectly classified, obtaining a performance equals to 100%. As shown in figure 4, minaret images class is one with best performance. This results can be due to the fact that all images in this class are rotated clockwise, which means that black triangles will be formed on top-left and bottom-right corners of the image (See figure 6). Additionally, the shape between minaret images are very similar, and share similar color intensities, for example, objects in figures 6(a) and 6(d).

On the other hand, Fig 9 shows the algorithm performed for test set of ImageNet dataset. As can see in Fig 9, 45 of the 200 categories of the test set have a low performance especially the Bedligton Terrier category whose images have a great variety since in some cases individual dogs are presented but in others they meet their owners (See figure 8). This fact changes the shape of objects of interest, which make algorithm to confuse the image and classify it in different categories. Additionally, given challenging factor such as illumination, occlusion or different agents present on the image that does not belong to the class, proper to the database, it becomes more difficult for the algorithm to correct in the classification of a query image.

## 5. Conclusions

To sum up, ImageNet represents a more challenging dataset that Caltech101, this can be due to it is composed of more realistic images, taking into account important as-

pects of object recognition task such as illumination, occlusion, background clutter or the variation of the point of view. This means a more difficult task for the algorithm which was evidenced in the results, obtaining a poor performance evaluating this database.

In order to improve the results obtained, it is recommended to create a joint descriptor which combines color, texture and shape representations into one complex classifier by the weight coefficient of each features. This approach would provides a more complete algorithm that can discriminate objects in images taking into account multiple kind of features than only its shape. Finally, due to the problem addressed is a multi-category classification task, the training of a classifier such as Random Forests could be more efficient given the fact that takes into account multiple classes, in contrast to SVM, which is binary and not so adequate to evaluate this computer vision problem.

## References


[1] A. Bosch, A. Zisserman, and X. Muñoz. Image classification using random forests and ferns. *2007 IEEE 11th International Conference on Computer Vision*, pages 1–8, 2007.

[2] O. Chapelle, P. Haffner, and V. Vapnik. Support vector machines for histogram-based image classification. *IEEE transactions on neural networks*, 10 5:1055–64, 1999.

[3] A. G. Faheema and S. Rakshit. Feature selection using bag-of-visual-words representation. *2010 IEEE 2nd International Advance Computing Conference (IACC)*, pages 151–156, 2010.

[4] L. Fei-Fei, F. R., and P. Perona. Learning generative visual models from few training examples: an incremental bayesian approach tested on 101 object categories. *IEEE CVPR*, 2004.

[5] H. Gao, W. Chen, and L. Dou. Image classification based on support vector machine and the fusion of complementary features. *CoRR*, abs/1511.01706, 2015.

[6] J. Han, M. Kamber, and J. Pei. Data mining: Concepts and techniques, 3rd edition. 2011.

[7] S. A. Hussein, H. E. Naby, and A. A.-H. Youssif. Image multi-classification using phow features. 2016.

[8] ImageNet. Overview. http://image-net.org/about-overview, 2016.

[9] W. Jian, Y. Huan, L. Jing, and X. Ping. Phow based feature detection for head pose estimation. In *2015 IEEE 16th International Conference on Communication Technology (ICCT)*, pages 437–440, Oct 2015.

[10] S. T. Klein and D. Shapira. Compressed matching for feature vectors. *Theor. Comput. Sci.*, 638:52–62, 2016.

[11] D. L. and Q. Weng. A survey of image classification methods and techniques for improving classification performance. 2007.

[12] X. Li, C. G. M. Snoek, and M. Worring. Learning social tag relevance by neighbor voting. *IEEE Transactions on Multimedia*, 11:1310–1322, 2009.

[13] Y. Liu, S. Liu, and Z. Wang. Multi-focus image fusion with dense sift. *Information Fusion*, 23:139 – 155, 2015.

[14] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60:91–110, 2004.

[15] D. L. Olson and D. Delen. Advanced data mining techniques. 2008.

[16] T. Shukla, N. Mishra, and S. Sharma. Automatic image annotation using surf features. 2013.

[17] M. J. Swain and D. H. Ballard. Indexing via color histograms. In *[1990] Proceedings Third International Conference on Computer Vision*, pages 390–393, Dec 1990.

[18] M. Toews and W. M. Wells. Sift-rank: Ordinal description for invariant feature correspondence. *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 172–177, 2009.

[19] M. Varma and A. Zisserman. Texture classification: Are filter banks necessary? In *CVPR*, 2003.


## Images

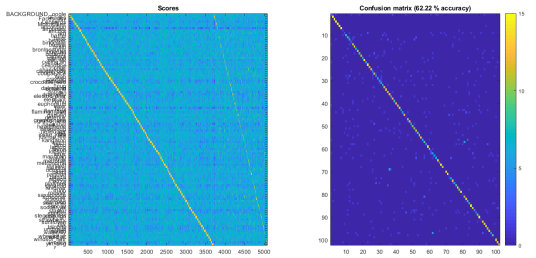



Figure 3. From left to right: Scores matrix and confusion matrix for Caltech dataset. ACA performance of PHOW method on this images is 62.22 %

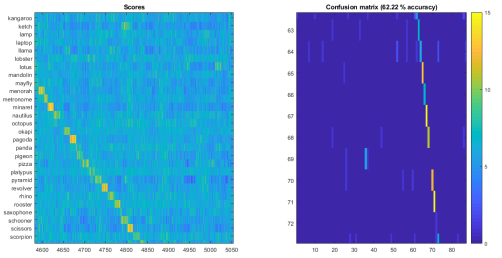



Figure 4. From left to right: Scores matrix and confusion matrix for Caltech dataset. ACA performance of PHOW method on this images is 62.22 % . Zoom to identify "hardest" classes

(a) Octopus example 1 (b) Octopus example 2 (c) Octopus example 3 (d) Octopus example 4

Figure 5. Example of images for octopus image classes observed in caltech 101 dataset.

(a) Minaret example 1 (b) Minaret example 2 (c) Minaret example 3 (d) Minaret example 4

Figure 6. Example of images for minaret image classes observed in caltech 101 dataset.

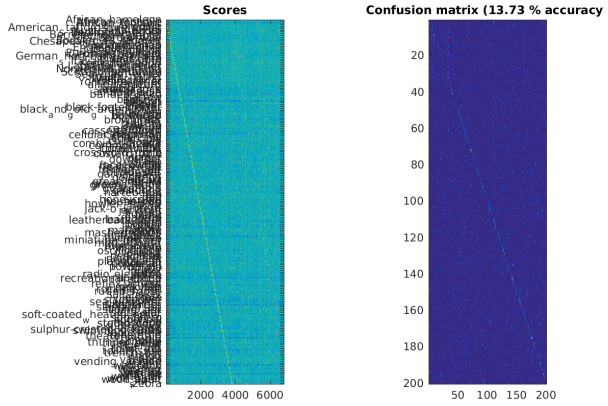



Figure 7. From left to right: Scores matrix and confusion matrix for train set of ImageNet dataset. ACA performance of PHOW method on this images is 13.73 %

(a) Bedlington terrier example 1 (b) Bedlington terrier example 1

Figure 8. Example of images for Bedlington terrier image classes observed in ImageNet dataset.

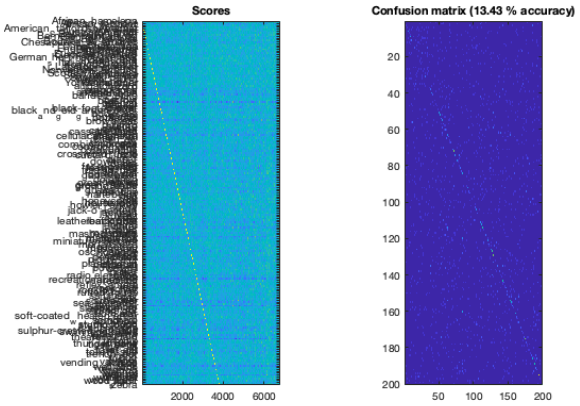



Figure 9. From left to right: Scores matrix and confusion matrix for test set of ImageNet dataset. ACA performance of PHOW method on this images is 13.43%