2017

Bag of words

COMPUTER VISION – LAB 11

JOSEP FAMADAS ALSAMORA / JORDI RIU VICENTE

UNIVERSITAT DE BARCELONA

UB | Facultat de Matemàtiques i Informàtica

Table of contents

Table of Figures

Table of Tables

1. Object recognition by BOW

a) Caltech101 database

We are given a function that using the PHOW descriptors technique tries to guess the class of different images, this technique will be explained throw the whole document. In *Figure 1* we can see an example performed with 10 different classes.

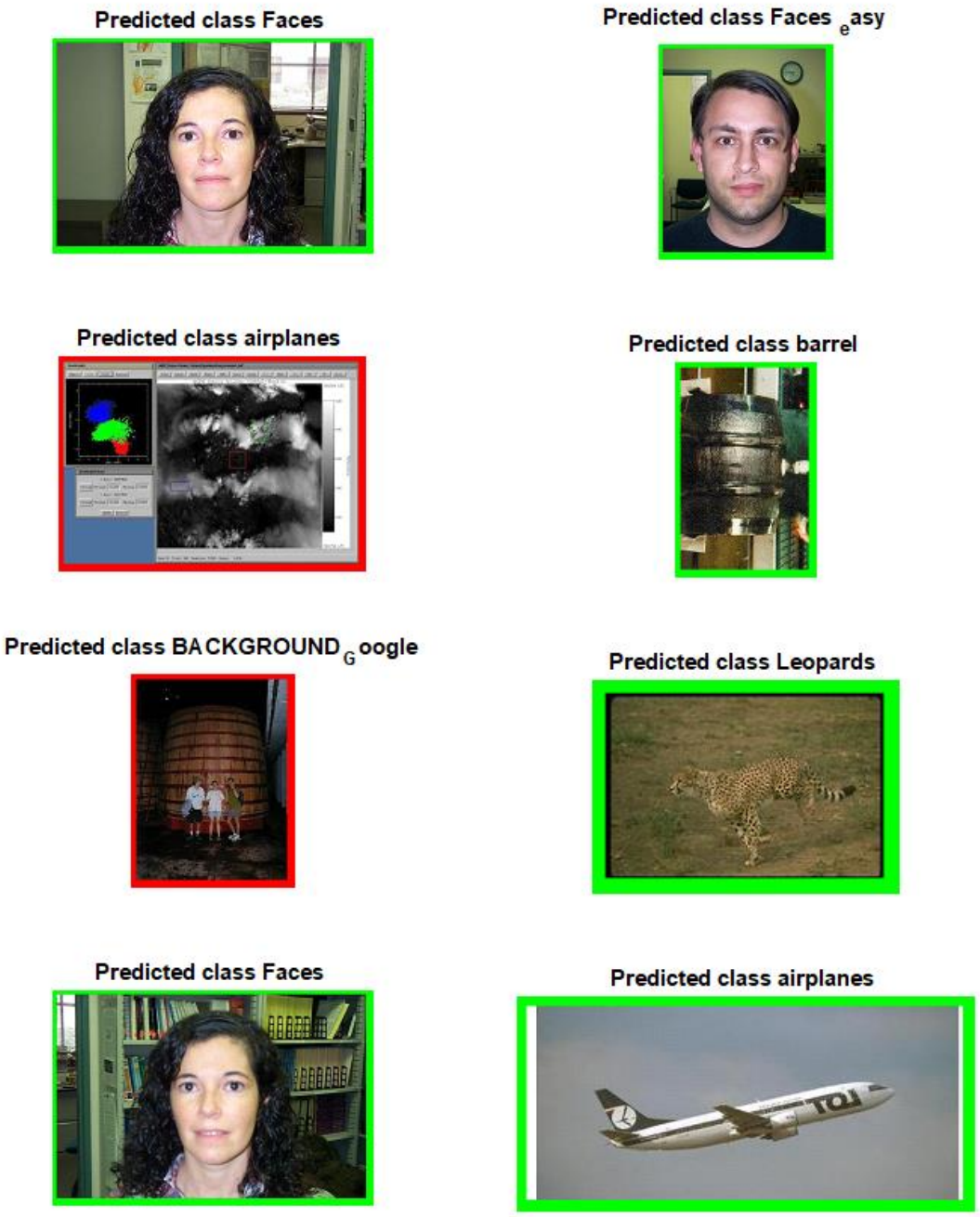

Figure 1: Example of the algorithm. In green the correctly labeled and in red the incorrectly.

In order to do the experimentation of this document we have created a new function called 'myphow_caltech101' which needs to be executed in the path 'vlfeat-0.9.20\apps'. It just needs to be executed using the green play button in MATLAB.

After one execution, all the files with the form 'tiny-something' in the path 'vlfeat-0.9.20\apps\data' must be erased.

b) PHOW descriptor

A PHOW descriptor is a combination of SIFT descriptors with a fixed scale and distance between them, forming a kind of dense grid. This is seen with two parameters:

- **Sizes:** Size of the SIFT descriptors of our PHOW. If we increment it, the number of SIFT descriptors is reduced, so if the parameter I incremented too much it can lead to an underfitting.
- **Step:** Distance between two consecutives SIFT descriptors. If we increment it, the result is similar to incrementing the Sizes.

c) Words

The words are the different visual features (such as wheel, door, eye, mouth…) in the images. They are represented by the centroids of the clusters and are generated during the training using the K-Means.

As said, they are the centroids of clusters of SIFT descriptors, so their dimension is the same of one of these descriptors, which is 128.

Too many words might lead to an overfitting and too few to an underfitting.

d) Vocabulary construction by K-Means

The input parameters of the K-Means are (sorted):

- The matrix containing all the SIFT descriptors.
- The number of clusters (words) we want.
- How is the K-Means applied. Elkan's algorithm in this case, which uses the triangle inequality to accelerate the K-Means.
- Maximum number of iterations, which prevents the system from getting stuck.

e) Spatial Histogram

The spatial histogram consists of dividing the image into small partitions and calculate the histogram of SIFT's in each of them, preserving the spatial information of each one of the histograms.

The dimension of the spatial histogram of an image is the number of sub histograms (which is the number of horizontal partitions of the image multiplied by the number of vertical partitions) multiplied by the size of a single histogram which is 300 (the number of visual words).

If we increment model.numSpatialX from 2 to 4, the dimension of the spatial histogram is also doubled.

f) SVM Classification

After the training process we obtain the weights and biases of each of the hyperplanes of our SVM classifier. We apply the one versus all method.

- dimension (W) = 3600x5
- dimension (b) = 1x5

For each image, we obtain the feature map of size 3600x1 and multiply it by the W transposed. Finally, we add the bias (b) and get the final scores for each class.

g) Image variations

Due to the fact that the orientation of the SIFT descriptors is constant, if we rotate the images, the results change.

In the same way, the size of the SIFT descriptors is also constant, so if we resize the images, the results change.

h) F-Score (1)

As we expected, in *Table 1* can be checked that if the number of categories is increased, the performance of the algorithm decreases, which can be seen by its F-score.

Number of categories	*F-Score*
5	0.9180
10	0.7978
15	0.7445
100	0.5580

Table 1: F-Score with different number of categories.

i) F-Score (2)

In *Table 2* we have computed the F-Score of the algorithm starting from some initial parameters and varying one of them each time.

```
conf.numClasses = 10 ;
conf.numSpatialX = 2 ;
conf.numSpatialY = 2 ;
conf.numWords = 300 ;
conf.phowOpts = {'Verbose', 2, 'Sizes', 7, 'Step', 5} ;
```

Figure 2: Initial parameters

Parameter Variation	*F-Score*
None	0.7978
numSpatialX = 1 || numSpatialY = 1	0.7316
numSpatialX = 4 || numSpatialY = 4	0.8288
numWords = 150	0.7838
numWords = 600	0.8655
Sizes = 5	0.8237
Sizes = 9	0.8047
Step = 3	0.7956
Step = 7	0.8162

Table 2: F-Score varying different parameters