Pascal project
Scene Segmentation and Interpretation

Jose Bernal Moyano[1] - Èric Pairet Artau[2]

[1] *jose.bernal@correounivalle.edu.co*
[2] *ericpairet@gmail.com*

May 29th, 2016

1 Introduction

Classification is one of the most challenging problems to address in the field of Computer Vision [1]. The goal is to find the most suitable label for a given input image according to its content and the knowledge about the different classes to which it may belong. Under this perspective, key questions should be analysed in advance:

- How to describe the objects of interest: finding the most suitable description for the considered objects is essential in the classification process. Since the classes may present variability caused by noise, viewpoint changes, illumination conditions, among other factors, the description has to be robust. Furthermore, the descriptors should not only condense the information of the objects but also be discriminant enough regarding the other classes.
- How to learn from the description of the objects of interest: once the descriptors are extracted for the different classes, the next step is to mine this information to obtain knowledge out of it. The approach consists of taking the set of features and finding the relevant information distinguishing the different classes; one from the others.

These two questions have been addressed by different authors in the field and, as a consequence, there are several options available for performing classification nowadays. Since all of them exhibit varied strengths and weaknesses, competitions such as Caltech101 [2] and The Pascal Visual Object Classes Challenge 2006 (VOC2006) are proposed. In particular, the aim of VOC2006 is to recognise ten objects classes (bicycle, bus, car, motorbike, cat, cow, dog, horse, sheep, and person) distributed in 1048 realistic scenes. Besides the fact that they are not pre-segmented images, the objects are placed in different weather conditions and, in some of them, more than one class occur.

In this paper, we propose several approaches for performing classification under the VOC2006 evaluation scheme. The paper is organised as follows. The strategy for tackling the classification problem is detailed in Section 2. The cornerstones of the proposed implementation are presented in Section 3. Different experiments were contemplated to test the classification proposal. The obtained results for each one of them are described and analysed in Section 4 and the best results are shown in Section 5. The project has been managed considering different tasks and expected times for solving them, which are detailed in Section 6. Finally, remarks of the project are exposed in Section 7.

2 Strategy analysis

Analysing the problem to solve before implementing a solution is an essential step. In this section, all the prior analysis performed to come up with a strategy able to approach VOC2006 is explained. Specifically, the constraints imposed for the implementation are studied in Section 2.1 while Section 2.2 details the proposed approach.

2.1 Implementation requirements

The project statement presents some requirements regarding the use of specific tools for the implementation. Thus, they have been analysed in this section to use them in an appropriate way. The general framework that had to be used as a basis for the implementation is detailed in Section 2.1.1. Then, the well-known SIFT descriptor and the structure where all descriptors have to be embedded in are introduced in Section 2.1.2 and Section 2.1.3, respectively.

2.1.1 Framework

Since the whole implementation was asked to be done in the VOC2006 framework, getting familiar to it was a primordial step. In short, such framework is structured into four folders, the content of them is introduced below.

- *local:* stores the features extracted from the training, test and validation sets.
- *results:* saves the documents containing the confidence of the classification for each class.
- *VOC2006:* contains the raw data to be used in the challenge, which is divided into three sets: test, train and validation. Only the first two are indispensable.
    - *Annotations:* consists of a set of annotations regarding the bounding boxes, the view, the image quality and the difficulty to obtain the correct classification.
    - *ImageSets:* has the ground truth for each class.
    - *PNGImages*: contains 5304 images related to the different classes. For speeding up the testing, only 1048 of them have been taken into account.
- *VOCcode:* holds a set of utilities provided by the competition to gather information related to the annotations, and display the Receiver Operating Characteristic (ROC) and compute the Area Under the Curve (AUC) associated with the classification results.

To keep the code related to our implementation separately from the basic framework, an extra folder called *src* was set up at the root of the framework. The external libraries used for the implementation, if any, are placed in this folder too.

2.1.2 SIFT

Until the beginning of this century, the comparison between images was carried out using a template matching approach in which a patch from one image was taken and compared to all the possible ones in another through similarity measures – for instance, Sum of Squared

Differences (SSD), Sum of Absolute Differences (SAD), Normalised Cross-Correlation (NCC), Zero-Normalised Cross-Correlation (ZNNC) [3, 4] and Median Correlation (MC) [5]. These kind of pixel-wise operations were able to deal with changes in illumination and also translation, but exhaustive searches were considered in order to be robust against scaling and rotation. As a result, the methods were very time-consuming.

One way to reduce the cardinality of the problem and, at the same time, the computational complexity associated with the process is to consider a subset of distinctive, complete and compact elements describing, without losing relevant information, the patch. A very well-known technique for detecting keypoints as well as describing them is called Scale-Invariant Feature Transform (SIFT) [6].

SIFT is a powerful technique since, as claimed by D. Lowe, which is able to achieve invariance against rotation, translation, scale and robustness against changes of viewpoint and illumination. Briefly described, the process consists of four main steps:

a. Scale-space extrema detection: in which the scale-space is computed and the initial keypoints are detected taking into account local minimum and maximum on the pyramid.

b. Keypoint localization: in which the candidates are refined based on their stability. Once the best ones are selected, their location and scale are calculated.

c. Orientation assignment: consists in looking for the highest peaks (i.e. greater than 80%) in the histogram of oriented gradients computed out for each keypoint. Using this approximation, the method achieves rotation invariance.

d. Keypoint descriptor: in which the feature is locally described using a histogram of gradients keeping in mind the location as well as the scale in which it was found.

In this paper, we use the SIFT detector and descriptor, but not its matching strategy.

2.1.3 Bag-of-Words

One of the key questions to solve when developing a strategy for classification is: which representation of the images should be used such that it is descriptive about their content and robust against different kind of alterations they may have. A way to overcome this problem is to use a well-known technique which has been applied successfully in text and also image classification, called Bag-of-Words (BoW).

The BoW representation appears originally in the field of document retrieval and is later implemented for object classification by translating the main components of the overall process [7]. The pipeline to follow under this perspective is illustrated in Fig. 1 and detailed as follows:

a. Feature extraction: different features are obtained from the training images using either a dense or sparse extraction approach.

b. Encoding: this step considers the transformation of the features into *Bag-of-Words.* Initially, the extracted features are clustered using, for instance, the K-means algorithm. Then, the resulting centres of the clusters, also referred as *words*, are used as terms of the *dictionary.* Afterwards, for each training image, the *Bag-of-Words* is computed by counting the occurrences it has of each cluster. Since the clusters do not exactly match the features of the image, the one with the highest similarity is chosen.

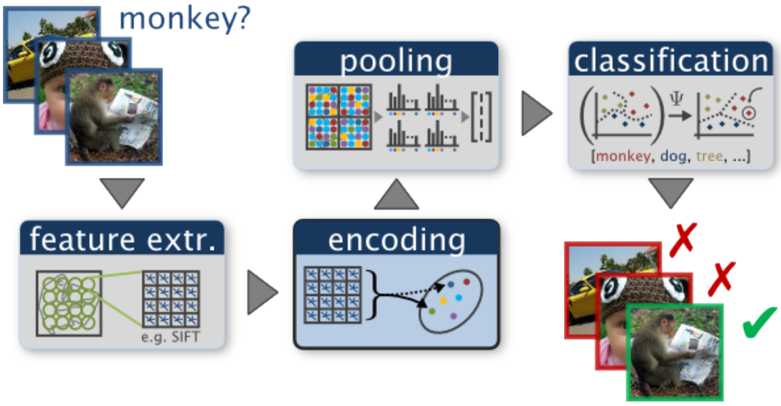


Figure 1: Classification through a BoW approach. Image taken from `www.robots.ox.ac.uk/~vgg/research/encoding_eval/`.

c. Pooling: if several spatial regions are defined, the pooling step consists in stacking the different bags to form the final descriptor of the image.

d. Classification: this step can be seen in two parts: training and testing. In the case of the former, the classifier is trained using the final bags and the label of each image. In the case of the latter, for every test image, the extraction, encoding and pooling are carried out and, finally, the classifier assigns the corresponding label.

2.2 Hypothesis of the classes

Before starting testing combinations of different descriptors, the first duty was to analyse the images in order to spot relevant information which could be useful to enhance the classification process. In this sort of ideas, we decided to group the classes into three categories: means of transport, animals and persons. Samples of each of the mentioned types were chosen and analysed to found key features. The resulting hypothesis are described in Section 2.2.1, Section 2.2.2 and Section 2.2.3, respectively.

2.2.1 Means of transport

The group means of transport contains the classes bicycle, bus, car and motorcycle. Some examples of this group are shown in Fig. 2. It can be seen that the objects belonging to this group are structured, i.e. they can be decomposed into basic geometrical shapes, such as circles, squares and lines. Moreover, the environment in which they are found contains several of these

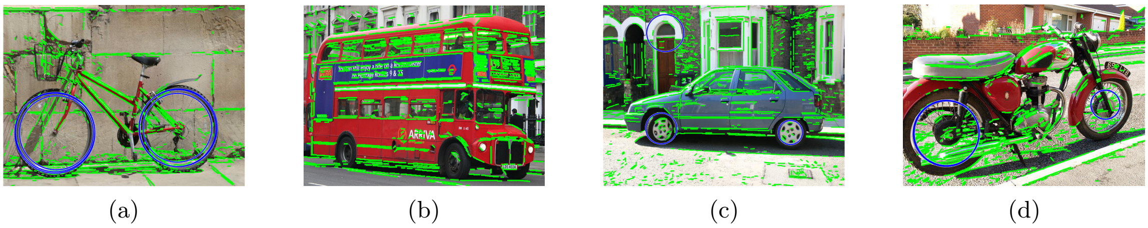


Figure 2: Geometrical shapes on means of transport extracted from the VOC2006 database.

forms. This fact can be advantageous to distinguish such group from the rest but not for an intra-group differentiation. Therefore, apart from identifying these figures using, for instance, the Hough transform, special features for each class should be gathered using SIFT.

2.2.2 Animals

Cats, dogs, horses, sheep and cows belong to the group of animals in the VOC2006 database. Two sub-groups can be found within this category regarding the places in which they are found: in indoor/outdoor environments or strictly outdoor scenarios. The hypothesis of both sets are described in Section 2.2.2.1 and Section 2.2.2.2, respectively.

2.2.2.1 Indoor/outdoor animals

Cats and dogs belong to this sub-group since they are found in indoor and outdoor scenarios as presented in Fig. 3. As a consequence, taking into account the surroundings of the object of interest is not going to give much information about the class itself. Thus, other techniques such as texture (e.g. obtained through a Gray-Level Co-occurrence Matrix (GLCM)), may help in the classification.

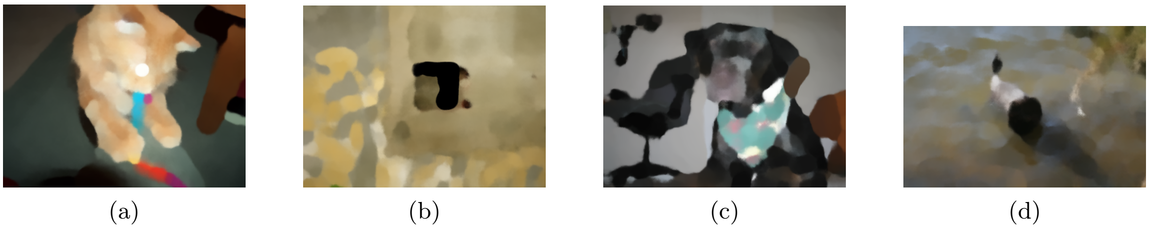

(a) (b) (c) (d)

Figure 3: Examples of indoor/outdoor animals extracted from the VOC2006 database with the respective texture.

2.2.2.2 Strictly outdoor animals

Cows, horses and sheep are the animals belonging to this sub-group. Unlike the previous ones, they are mostly found in outdoor environments as it can be seen in Fig. 4. Therefore, a strategy to classify animals belonging to this group is to use information of the environment, such as grass or sky. Also, these animals evidence particular textures which can be used to differentiate them from the other classes. These two aspects can be achieved by considering information from the colour spaces and texture.

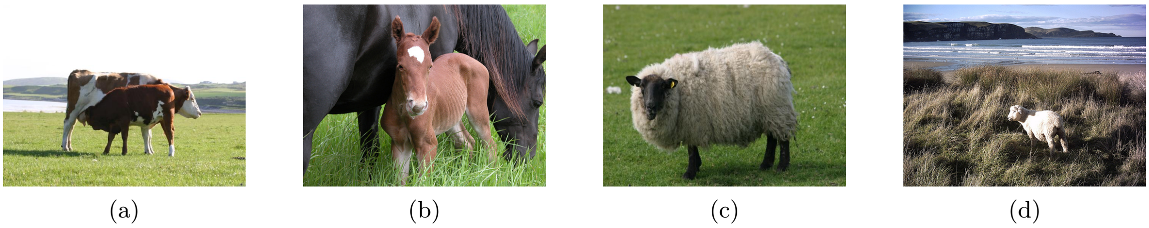

(a) (b) (c) (d)

Figure 4: Examples of outdoor animals extracted from the VOC2006 database.

2.2.3 Persons

After looking at the results of VOC2006, persons is expected to be one of the most complicated classes to classify. This fact takes place due to the wide variability of clothes appearing in the images of this class and the low amount of representative features of persons. Thus, as shown in Fig. 5, detection of skin and face on this category could be essential to achieve good results. Moreover, since persons can be found in different scenarios, it is better to avoid using a considerable amount of contextual information.

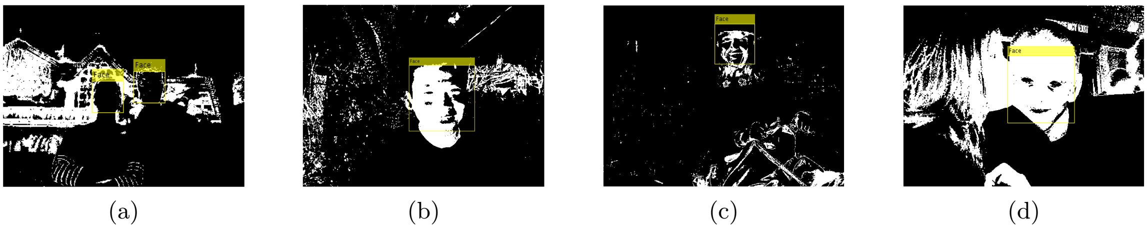


Figure 5: Examples of persons extracted from the VOC2006 database with the corresponding label over their faces and skin detection.

3 Implementation

The integration of the strategy introduced in Section 2 into the VOC2006 framework is detailed in this section. Specifically, the process contemplates four steps: extraction and concatenation of features, training of the classifier and testing the full framework, which are respectively described in Section 3.1, Section 3.2, Section 3.3 and Section 3.4.

3.1 Features extraction

Features can be extracted either for the full image (global features) or from different patches of it (local features). This latter approach is sub-divided in the literature into two groups: dense and sparse. The features included in the first group are extracted regularly along the full image. In contrast, the sparse features are computed exclusively in specific locations of the image which exhibit relevant information.

The considered global, dense and sparse features are introduced in Section 3.1.1, Section 3.1.2 and Section 3.1.3, respectively. Regardless the approach, once the features are extracted from the images, they are stored into the *local* folder provided by the original framework to save some time.

3.1.1 Global

All features below have been considered for describing the images in a global way, i.e the information provided for each descriptor is encapsulated into a unique value.

- *Gr:* mean of the image red channel.
- *Gg:* mean of the image green channel.

- *Gb:* mean of the image blue channel.
- *Gh:* mean of the image hue channel.
- *Gs:* mean of the image saturation channel.
- *Gv:* mean of the image value channel.
- *Ggeom:* being c the number of circles and l the number of long lines detected in the scene, the score given by this feature corresponds to

  $$Ggeom = w_1 \cdot c + w_2 \cdot l, \tag{1}$$

  where w_1 and w_2 are the weights of both features.
- *Gface:* combination of Viola-Jones algorithm and skin detection provided by MATLAB. Due to the high number of false positives provided by the face detection algorithms, the final verdict is subjected to the criteria of skin detection. If both agree, the value of the Gface depends on the area of intersection between both methods.

3.1.2 Dense

The extraction of features in a dense way consists in taking the information of different pixel neighbourhoods along all the image. Specifically, such patches were set up with a size of 16×16 pixels, and a spacing between them of 10 pixels.

- *Dsft:* SIFT descriptor computed in a dense way along the grayscale version of an image.
- *Dsft-h:* SIFT descriptor computed in a dense way along the image hue channel.
- *Dsft-s:* SIFT descriptor computed in a dense way along the image saturation channel.
- *Dsft-v:* SIFT descriptor computed in a dense way along the image value channel.
- *Dhog:* HOG descriptor computed in a dense way along the original image.
- *Dtext:* texture descriptor computed in a dense way along the grayscale version of an image.

All the SIFT-based and HOG-based descriptors have been computed using functionalities provided by the VLFeat library.

3.1.3 Sparse

In contrast to the previous strategy, this one considers extracting the information of the most distinguishable pixel neighbourhoods. For this purpose, the SIFT detector provided by the VLFeat library has been used to detect the keypoints. Then, the features introduced below were computed at each keypoint.

- *Ssft:* SIFT descriptor computed in a sparse way along the grayscale version of an image.
- *Ssft-h:* SIFT descriptor computed in a sparse way along the image hue channel.

- *Ssft-s:* SIFT descriptor computed in a sparse way along the image saturation channel.
- *Ssft-v:* SIFT descriptor computed in a sparse way along the image value channel.
- *Shog:* HOG descriptor computed in a sparse way along the original image.
- *Stext:* texture descriptor computed in a sparse way along the grayscale version of an image.

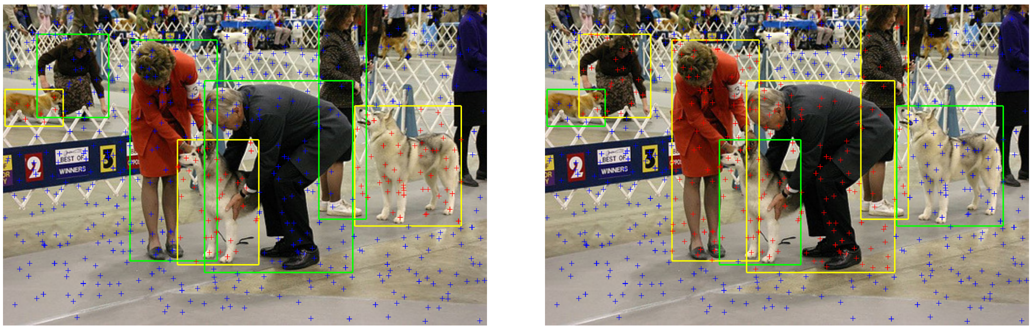
As mentioned in the previous section, all the SIFT-based and HOG-based descriptors have been computed using the VLFeat library.

3.2 Features concatenation

The VOC development kit can be used to obtain useful information regarding the training set, such as the part of the image which really contains a specific class. As illustrated in Fig. 6, the objects are located withing bounding boxes. As mentioned in Section 2.2, the environment may give important information (i.e. context) to some classes, while in other cases it may be better to avoid it. Taking this into account, we divided the features into three sets:

- iBB: referring to the features within the bounding box.
- oBB: referring to the features outside of the bounding box.
- non: referring to the features not belonging to the class at all.

Thus, the strategy is to take a defined number of samples from the oBB set instead of considering it completely, if necessary.



(a) Class of interest: dog

(b) Class of interest: person

Figure 6: Differentiation of features according to the information they provide.

3.3 Classifier training

Once the features are extracted using the different approaches presented before, the next step is to train the classifier. Initially, the BoW for each training image is computed as presented in Section 2.1.3. Recall that if the strategy contemplates spatial divisions of the image as presented in Fig. 7, the BoWs will be appended, one after another. Afterwards, this information along with the corresponding labels are used to train the considered classifier. Specifically, it is based on a Support Vector Machine (SVM) provided by the MATLAB platform.

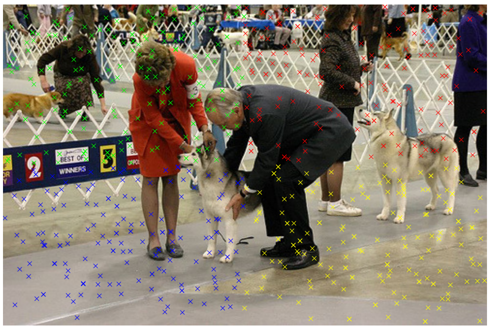


Figure 7: Differentiation of features according to their spatial information.

3.4 Strategy testing

The evaluation of the strategy is the final step of the process. Given an input image, its features are extracted, the corresponding BoW is computed and, then, the classifier is used to determine the most suitable label. The output of the implemented SVM classifier would be the corresponding label as well as the confidence in determining this verdict. The latter value is used subsequently to obtain the ROC through a piece of code provided in the VOC2006 framework.

4 Experimentation

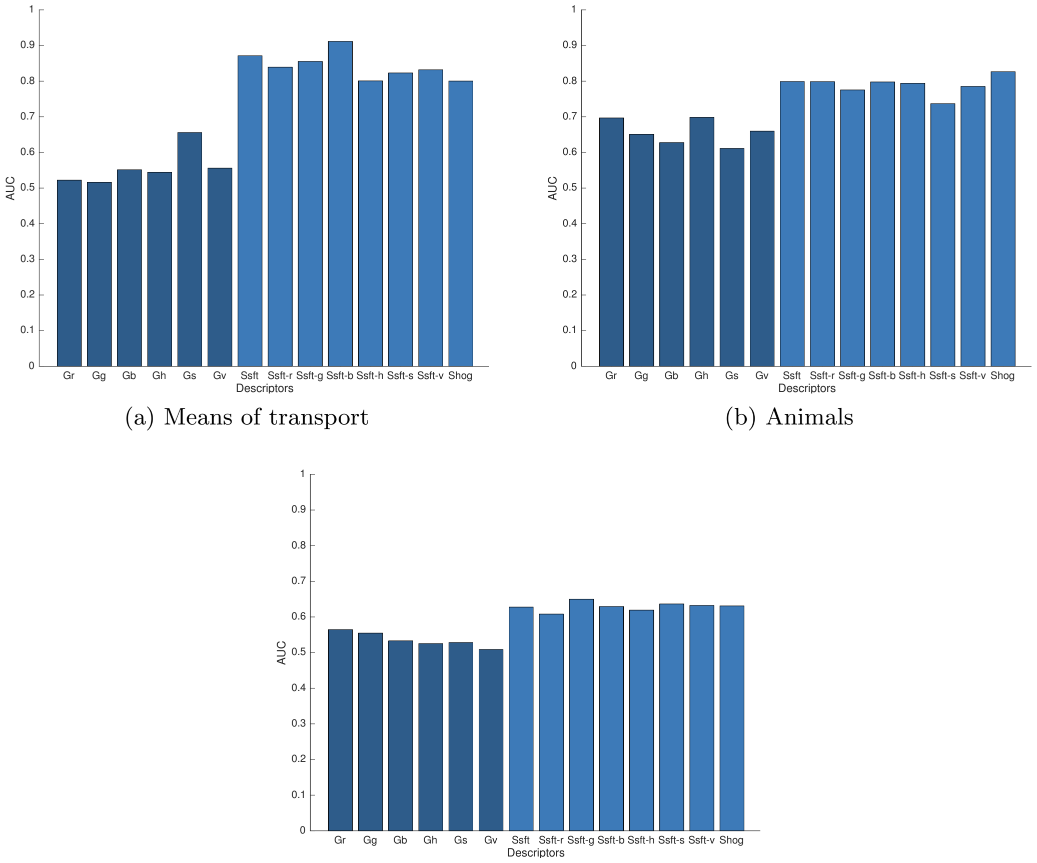
According to the hypothesis presented in Section 2.2, we designed a set of trials to be tested in the bicycle, cat and person classes, which are representative classes of the groups stated in the proposal, i.e. means of transport, animals and persons. Initially, the effectiveness of the descriptors implemented in Section 3.1 were analysed in each group, the results of which are shown in Section 4.1. Then, due to the structure of the considered framework, the following aspects wanted to be analysed too:

- Number of clusters: depending on the complexity (i.e. amount of details) of a class, the number of considered clusters might lead to a different classification. The results regarding this aspect are shown in Section 4.2.
- Features filtering: as mentioned in Section 2.2, considering the environment can be useful for some classes. The influence of considering one amount of features outside the bounding box or another is evaluated in Section 4.3.
- Spatial information: according to Section 3.3, the implemented framework can either consider spatial information or not. The influence of this approach is explained in Section 4.4.
- Training data augmentation: theoretically, the more information you provide to a machine learning system, the more reliable would be. Thus, it has been considered to increase the amount of data to train the classifier in order to improve the final results. The variation of the results when considering this approach it is reported in Section 4.4.

In order to evaluate the cross-influence of the features, the number of clusters and the filtering of features, 2464 trials were initially performed. The corresponding results have been visually summarised throughout different plots, which have been extremely helpful to extract valuable conclusions for the final experiments. Moreover, they have allowed bounding the dimensionality of such problem when testing both the spatial information and the augmentation of the data training.

4.1 Descriptor analysis

From the bar charts in Fig. 8, it can be concluded that the sparse descriptors are able to describe the images better than the global ones. This behaviour is observed since the basic global features take into account the mean in the different channels. Therefore, this information may not be enough to represent the content of the object of interest since it is combined with the one from the environment.



(a) Means of transport

(b) Animals

(c) Persons

Figure 8: Best results achieved for the features considered initially.

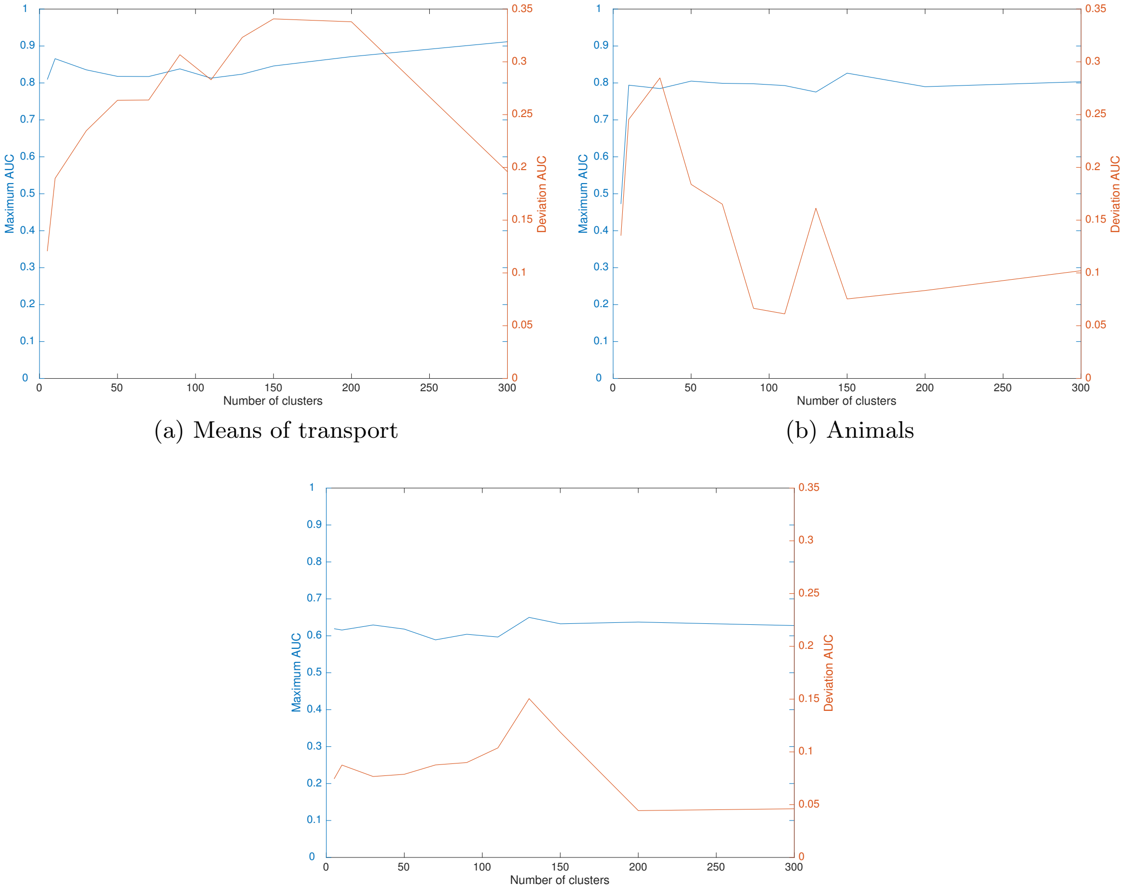
4.2 Cluster analysis

The number of clusters to select is one of the key factors determining the results of the classification. Thus, we decided to perform an extensive comparison to decide, initially, which should be its value. The obtained results are presented in Fig. 9, from where the following three highlights can be extracted:

- As shown in Fig. 9(a), the number of clusters can vary the final AUC up to 0.1.

- The higher the number of clusters, the lower the variability of the data. In any case, a stable classifier is preferred.
- If a unique amount of clusters has to be selected for all the classes, 300 would be the best option.

However, these conclusions are more probable to happen if the descriptors are considered alone (which is not the desired case). Thus, it is expected that the number of clusters to consider is, somehow, proportional to the cardinality of the data; the larger the length of the codeword, the higher the number of codewords.



(a) Means of transport

(b) Animals

(c) Persons

Figure 9: Maximum and deviation of the AUC value for different classes regarding the number of clusters.

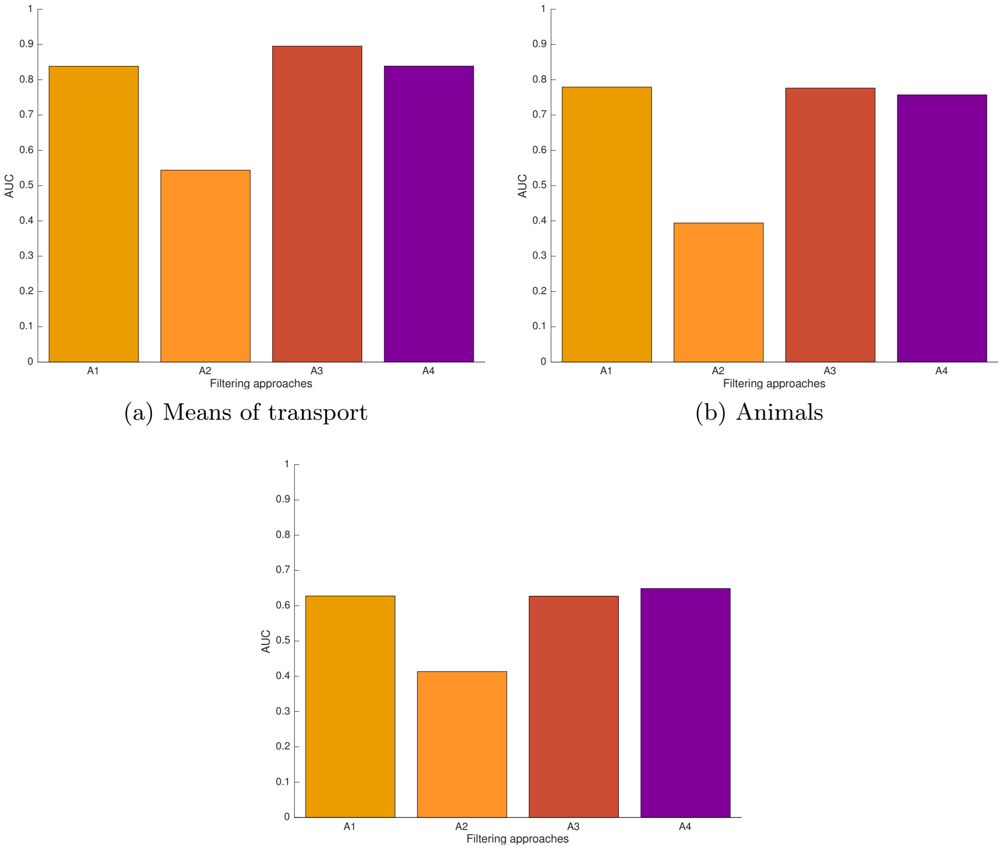
4.3 Data filtering analysis

For some classes, the environment is not a suitable option since it may not give contextual information while for some others considering it is essential. Four approaches, described in Table 1, were considered to evaluate the influence of including data from outside of the bounding box (referred in the table as *oBB*) and from the features obtained for images in which the class is not present (referred in the table as *non*).

Approach	Samples from *oBB*	Samples from *non*
1	100%	100%
2	50%	66%
3	50%	100%
4	0%	100%

Table 1: Number of samples from *oBB* and *non* for the considered approaches.

The results regarding the AUC are presented in Fig. 10. It can be observed that considering only a subset of the information from the negative samples is not an option since it reduces the final result. On the other hand, affecting the number of samples outside of the bounding box does not affect it considerably. Thus, the training should be performed taking into account only three of the four configurations.



(a) Means of transport

(b) Animals

(c) Persons

Figure 10: Maximum value obtained using different data filtering approaches. Note that A_i stands for the $i-th$ approach presented in Table 1.

4.4 Spatial information and training data augmentation

The influence on the result when the spatial information is considered or not has been empirically determined. Specifically, for each of the analysed classes, the parametrisation giving the best results was considered in a double trial to compare the influence of this feature of the framework.

The obtained results stated that considering spatial information is not always beneficial; when the objects are not in the spotlight of the scene, dividing the image in spatial frames corrupts the BoW content. According to it and after analysing some of the images, we concluded that further analysis of this option should be carried out to use it correctly.

Regarding the augmentation of the data to train the classifier, it was achieved by not only considering the training set of the VOC2006 framework, but also the validation one. After doing some trials in the different classes, we concluded that considering it is completely beneficial to enhance the results; the bigger the amount of information given to the classifier regarding a class, the higher the probability of classifying an image correctly.

5 Results

Taking into account the information obtained in the previous section about the different parameters of the framework, more experiments combining the different descriptors were performed. In this section, we detail the timing for performing each of the classification steps, the best results obtained for each class, and a discussion of them in Section 5.1, Section 5.2 and Section 5.3, respectively.

5.1 Framework analysis

The analysis of the framework consists of determining the time required for performing the classification of the images. The results of the feature extraction step are presented in Fig. 11. It can be seen that *Ggeom* and *Gtext* are the descriptors taking more than three times to be computed in comparison to the others. This is actually expected since, for both cases, circles, lines and faces are detected.

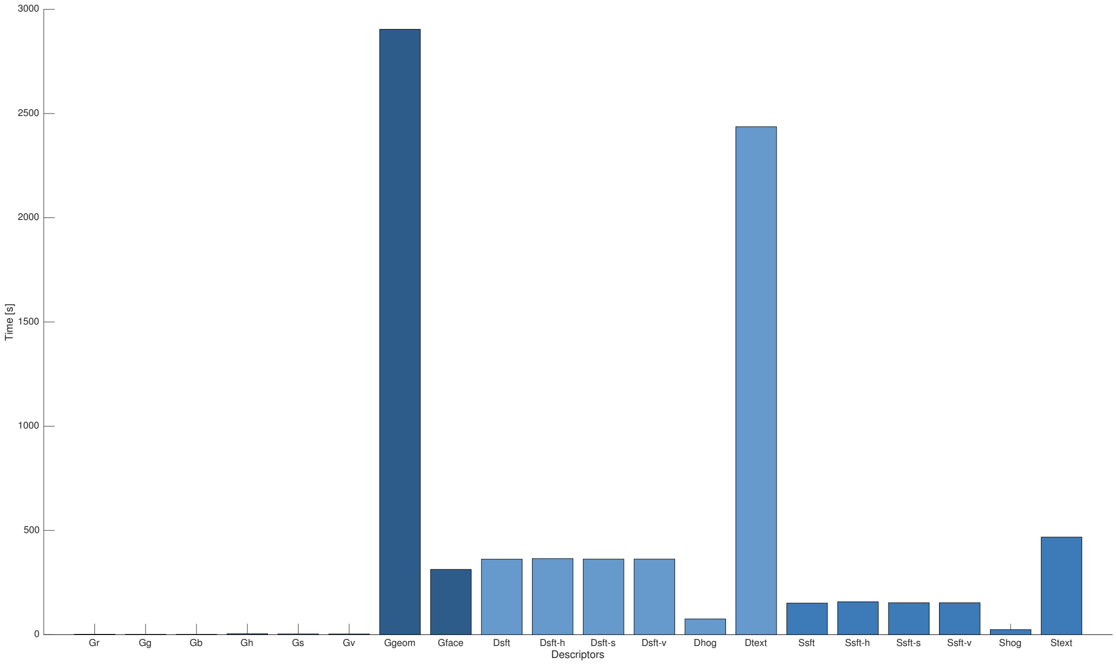
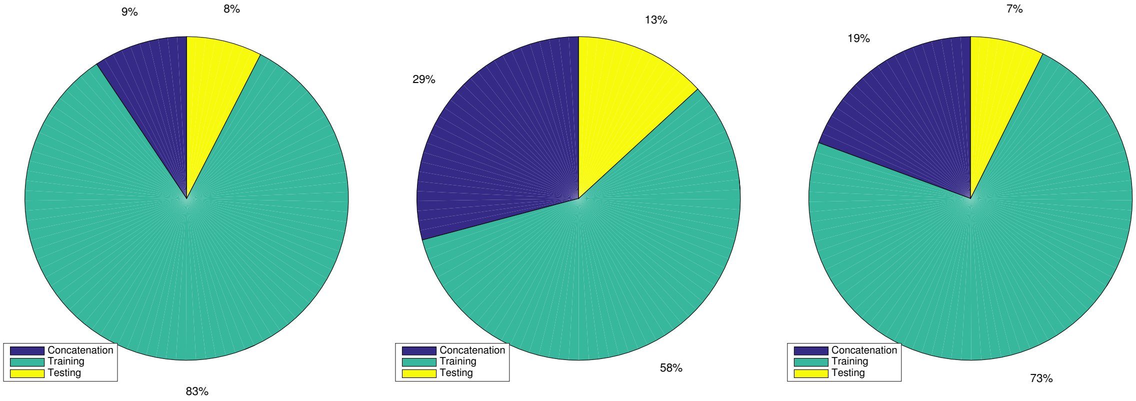


Figure 11: Time spent extracting the considered features in the 1048 images.

In the different trials that were evaluated, the timing for performing the different processes varied completely depending the number of clusters and the complexity of calculating the features, among other factors. Thus, we selected the worst scenarios to give an idea of the timings. The results are shown in Fig. 12. It can be observed that the training part was the most consuming step since the clustering is carried out over hundreds of thousands of features. In contrast, the testing part is the one taking less time to compute.



(a) Dense SIFT in H, S and V (b) Dense HOG (c) Dense SIFT in grayscale

Figure 12: Time spent for three steps of the framework regarding three different descriptors.

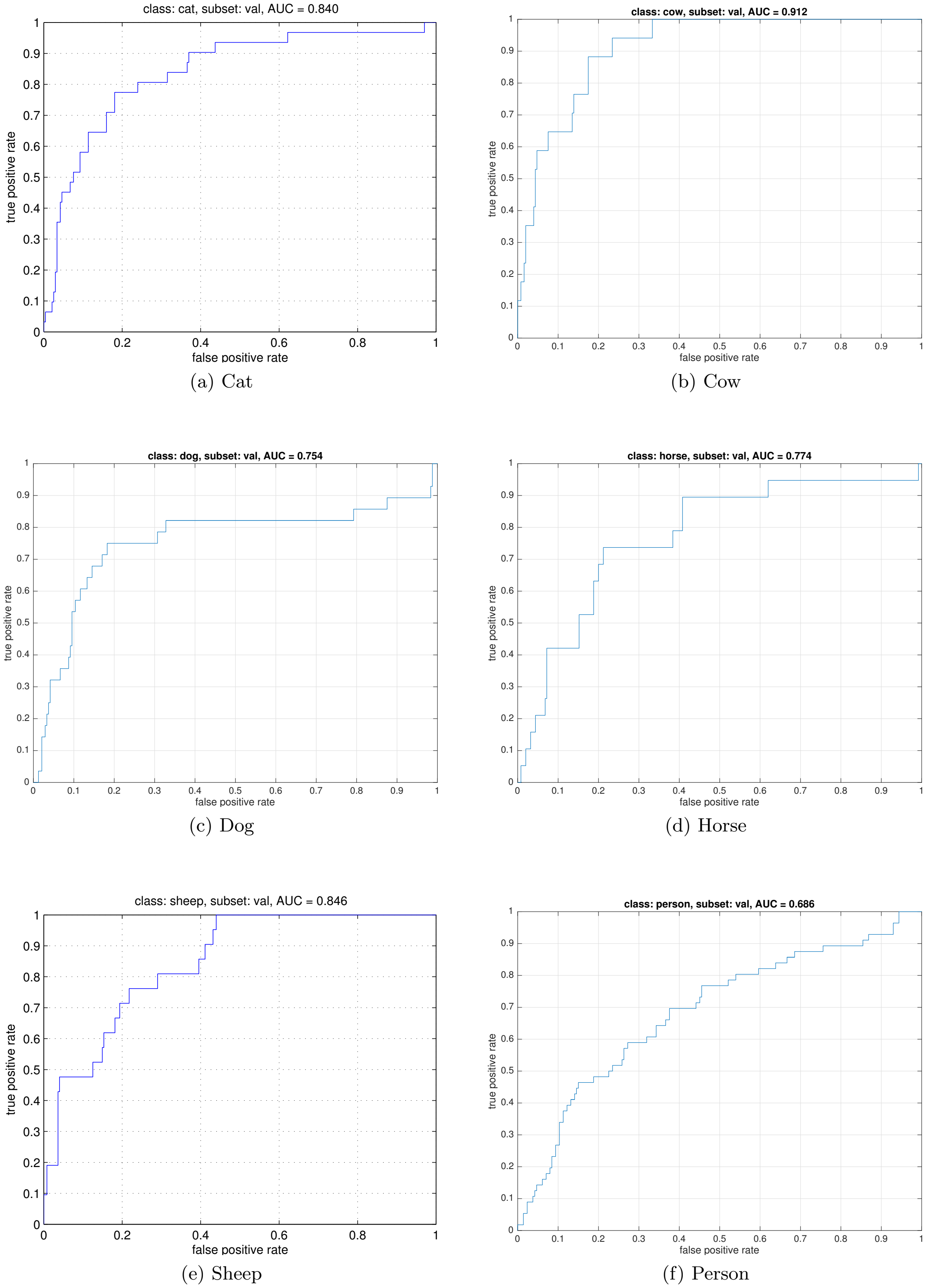
5.2 Best results

The best results obtained for each class are presented in Fig. 13 and Fig. 14, and summarised in Table 2. It can be observed that the approaches were able to correctly identify categories such as bicycle, cow and motorbike, since it achieved an AUC value greater than 0.90, while for dog, horse and person the AUC value did not surpass 0.80. In average, the obtained AUC score of our implementation was of 0.84.

Regarding the hypothesis formulated initially, it can be seen that (i) descriptors calculated in a dense way required more clusters to achieve good results and (ii) the best results were obtained for the group of means of transport using only 0.25 of features outside of the bounding boxes. Regarding the category of animal, the best results were obtained considering all the features describing their environment.

Class	**AUC**	**# clusters**	**Features**	**oBB rate**
Bicycle	0.92	100	*Ggeom*, *Ssft*, *Ssft-h*	0.25
Bus	0.84	200	*Ggeom*, *Ssft-h*	0.25
Car	0.89	800	*Dhog*	0.25
Cat	0.84	200	*Ssft*, *Ssft-h*, *Stext*	1.00
Cow	0.91	300	*Gg*, *Ssft*	1.00
Dog	0.76	800	*Dhog*	0.25
Horse	0.77	100	*Shog*, *Stext*	1.00
Motorbike	0.94	800	*Dsift-h*	0.25
Person	0.69	70	*Gface*, *Shog*, *Stext*	0.00
Sheep	0.85	300	*Ssft*, *Shog*, *Stext*	1.00

Table 2: Final results and their configurations for the ten classes proposed by VOC2006.



(a) Cat

(b) Cow

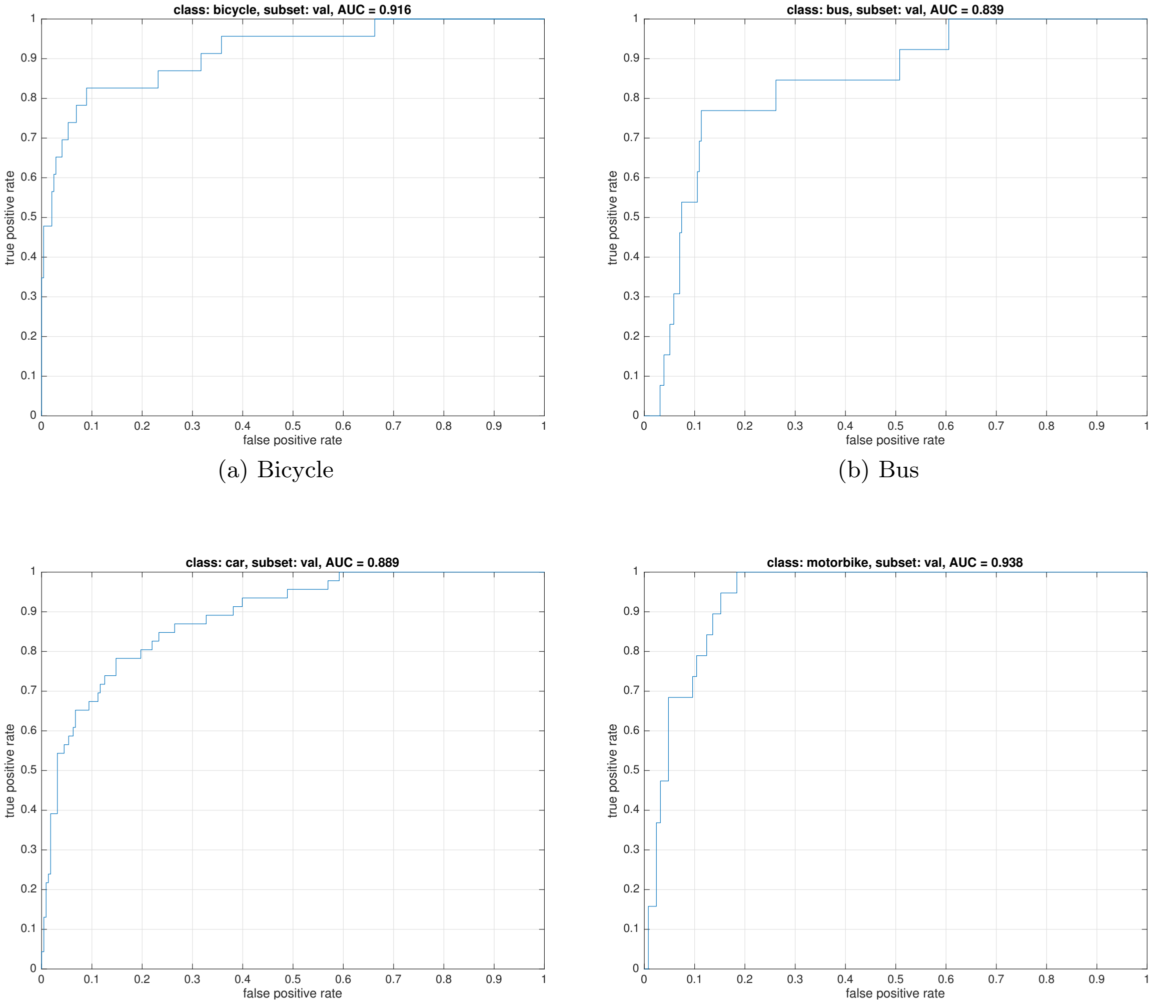
(c) Dog

(d) Horse

(e) Sheep

(f) Person

Figure 13: ROC and AUC for the best results achieved for the group animals and persons.



(a) Bicycle

(b) Bus

(c) Car

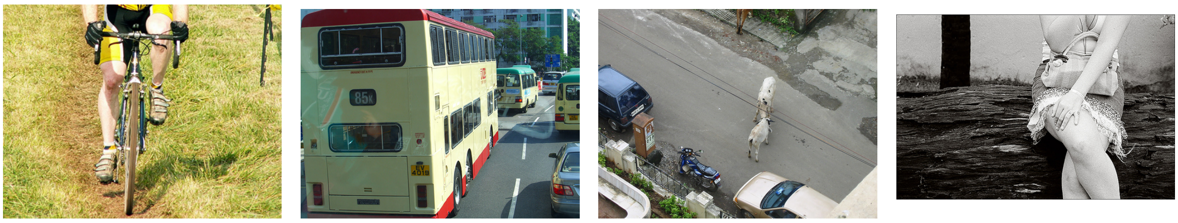
(d) Motorbike

Figure 14: ROC and AUC for the best results achieved for the group means of transportation.

5.3 Discussion

Taking into account the results provided in the previous section, it can be concluded that the assumed strategy had some limitations. Some of them are detailed as follows:

- It can be seen that the description based on geometrical figures is not able to detect the corresponding classes in Fig. 15(a) neither in Fig. 15(b) due to the perspective of the objects.
- The usage of features from the surroundings was not suitable in the case of Fig. 15(c) since it was assumed that cows were strictly in natural environments with grass and sky.
- In some cases such as the one presented in Fig. 15(d), it is not suitable because the face of the person is not shown and the image is in greyscale.
- Since HOG is based on gradients, illumination conditions may condition whether a match is found or not.



(a) (b) (c) (d)

Figure 15: Examples where the proposed approach might fail classifying the images.

Moreover, the whole process relies on the selection of good clusters obtained through the K-means. That means that a proper selection of the number of clusters is a key factor determining the classification outcome. We observed that this value depends, in a way, on the cardinality of the descriptor; the largest it is, the more the clusters to consider. However, to the best knowledge of the authors, there is no way to determine the clusters leading to outperforming results.

6 Project management

When working in a team, the project management takes special importance. A bad organisation may lead to unnecessary waste of time, efforts and, consequently, penalise the final work. This fact is why we have used two platforms to organise the tasks and synchronise the work.

A platform called Bitbucket was used to share the code, keep track of all the changes, notify the issues found during the implementation and assign tasks to the members of the group. Another one named Sharelatex was used to write down the final report in an efficient and collaborative way. This platform allows to all the members of the group to edit at the same time.

In general, the work done for this laboratory session has been divided into the following six tasks:

- Task 1: familiarisation with the provided framework.
- Task 2: analysis of the problem and design of a strategy.
- Task 3: initial implementation of the strategy into the framework.
- Task 4: experimentation and adjustments to the framework.
- Task 5: final tests of the strategy.
- Task 6: overall analysis and creation of the report.

In contrast to the organisation adopted in the laboratory sessions, for this project, we found more appropriate to work hand in hand with the other throughout all the project. This strategy makes us schedule the duration of the different tasks at the very beginning of the project to deliver all the work in time. Apart from this, the final amount of time needed for each task has also been computed in order to improve our planning in future works. All this information is condensed in Table 3.

Task number	**Expected duration (hours)**	**Real duration (hours)**
1	5	6
2	5	12
3	30	20
4	20	50
5	10	5
6	20	20

Table 3: Expected timing for the realisation of the different tasks.

Even though some of the implementations have been done individually by one or other member of the group, the peer-review process forced the other to understand, check, criticise and correct the others' code. Only in that way, an equivalent amount of work when it comes to implementation can be attributed to both members of the group. About the last task, which consists of writing the report, it was carried out simultaneous by the two members of the group.

7 Final remarks

In this project, a framework for classifying non-pre-segmented images was designed, analysed, implemented and evaluated under the VOC2006 framework. The strategies considered within the proposal, which contemplates state-of-the-art as well as *ad hoc* descriptors, were developed after a prior study of the image database.

Hypotheses regarding the different classes were formulated at the beginning of the project in order to achieve better classification results. Some of them were reinforced during the experimentation process, such as the amount of contextual information to consider, while some others were undermined by looking at the cases in which they were not working, such as the assumption that images of persons are always showing their faces.

The implementation of the proposal has been developed on top of the VOC2006 basic framework, which not only provides all the images and their corresponding information, but also some source code for loading the images, their ground truth and displaying the classification results. Once our approach was integrated, the final framework consists of four steps: feature extraction and concatenation, and then training and testing of the classifier.

Along this project, we focused our work on evaluating global and local descriptors gathered in sparse and dense ways, a number of codewords, and the inclusion of spatial and contextual information. Thus, an extensive evaluation of these parameters involved in the framework was performed to detect their influence on the classification process. Three important aspects were found in this assessment: (i) global descriptors are not as good as local descriptors for describing the content of an image, (ii) the number of clusters determines the variability of the final results, and (iii) the contextual information is a decisive factor in the classification process. These three facts were taken into account to perform the final experiments of the project.

The results showed that the proposed framework was able to obtain values of AUC greater than 0.84 for seven out of ten classes, three of them being greater than 0.91. Also, the classification highlighted drawbacks of some descriptors such as poor detection under viewpoint changes and occlusions. Although it was noted that the cardinality of the descriptor may influence the number of codewords to be considered, further research should be carried out to be conclusive with respect to this statement.

Finally, the authors encourage the readers to analyse different configurations using spatial information obtained not only in quadrants as it was presented in this paper, but also in vertical or horizontal divisions. Additionally, a proper study of the synergy (i.e. the compatibility) among the features should be considered to improve the final results.

References

[1] R. Fergus, P. Perona, and A. Zisserman. Object class recognition by unsupervised scale-invariant learning. In *In CVPR*, pages 264–271, 2003.

[2] L. Fei-Fei, R. Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. 2004.

[3] G. S. Cox. Template matching and measures of match in image processing, 1995.

[4] C. Sung-Hyuk. Comprehensive survey on distance/similarity measures between probability density functions. *International Journal of Mathematical Models and Methods in Applied Sciences*, 1(4):300–307, 2007.

[5] M. Trujillo and E. Izquierdo. A robust correlation measure for correspondence estimation. In *3D Data Processing, Visualization and Transmission, 2004. 3DPVT 2004. Proceedings. 2nd International Symposium on*, pages 155 – 162, sept. 2004.

[6] David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004.

[7] Yin Zhang, Rong Jin, and Zhi-Hua Zhou. Understanding bag-of-words model: a statistical framework. *International Journal of Machine Learning and Cybernetics*, 1(1):43–52, 2010.