University of Girona

Scene Segmentation and Interpretation

Final Project

Pascal Challenge

Author:
Di MENG
Darja Stoeva

Supervisor:
Xavier Lladó
Arnau Oliver

June 5, 2017

1 Introduction

This paper studies object classification using the method bag of words. Object classification aims to predict weather a specific class object is present in an image. It does so by first trianing a classifier with extracted features from a train data set of images with ground truth lables and then assigning a label to a new, test image. The most important steps in object classification are the feature detection, description and matching. Depending on the data set we are dealing with and the objects we want to detect, different feature detectors and describers can be used.

This project is based on the Pascal Challenge 2016. The main challenge is to recognize objects from a number of visual object classes in realistic scenes. In this project we have ten classes with hundreds of images. They are supposed to be classified into correct classes. Along with the image datasets, the annotation and ground truth are also provided for better classification and evaluation.

The difficulty of visual objects classification is that the objects we focus on in the images are with different scales, different orientations, different expressions. Also there are many unnecessary background information confusing the descriptors. In terms of the large dataset, the computation time is also a factor influencing the performance of the results since it is expensive to do the improvement and modification. There are several papers that analyse the performance and compare diffenerent ways of finding key points and different methods of extraxcting the features [2, 3, 4]. Our approach is ispired by one of the methods in [2] and is based on an SVM classifier trained on feature vectors built using local image descriptors. The follwoing sections explain more about the way we have approched the object classification, discussion of the results we have obtained and finally a conclusion.

2 Methodology

The Pascal data set contains 10 classes of the following objects: bicycle, car, sheep, dog, cat, person, cow, bus, horse and motorbike. The data is already splited into training, testing and evaluating sets, where each of them contains a possitive image, meaning the object from the class is present in the image, and negative image, meaning the object of the class is not present.

Our method of classification is based on the bag of word approach us-

ing Laplacian of Gaussian detector with SIFT descriptor and a linear SVM classifier. Firstly, we extract features from the training data set, then we construct the dictionary, from which we encode the feature histogram, or the bag of words, for each trainging image and finally we train the classifier with the collection of all feature histograms. In order to test the classifier, the features from the test data were extracted and encoded with the same principal as for the training. The classification task was performed several times with different parameters. The following subsection explain more about each of the steps.

2.1 Feature Detection and Description

In order to match two images, their feature vectors need to be obtained and compared. However, before computing the feature descriptors, a feature detection is performed to get the keypoints from which the descriptors will be extracted. For our approach we detected 60, 120, 350 and 1000 keypoints using a Laplacian of Gaussian, using a modified code inspired from [1]. Then to extract the descriptors of these keypoints SIFT descriptors are used from the VLFeat library.

2.2 Bag of Words

Once we have extracted the features from all training images from one class, we use Gaussian mixture model (GMM) to cluster the features into clusters which form the dictionary. The number of clusters was changed when the number of keypoints was changes and it can be seen in Table 2. Then, to encode the images into histogram vectors we use the Fisher encoding.

Table 1: Number of keypoints and clusters used

	Keypoints	Clusters
classifier 1	80	60
classifier 2	120	100
classifier 3	350	150
classifier 4	1000	300

2.3 Classification

The aim of this ste is to predict whether at least one object of a given class is present in an image. For this purpose, a svm classifier using the VLFeat libabry was trained with the encoded training set histograms for each class. Since the `vl_fisher` normalizes the histograms, if specified, the histograms did not needed to be normilized. Then the classifier was tested with the testing set using binary decision, one classifier per class. Again, each of the images in the training set was encoded using bag of words then mapped with the classifier to predict the object class.

3 Results

Figure 1, 2, 3 and 4 show the ROC curves obtained for each classifier when using different parameters. Figure **??** shows only the results of the first two classes since the MATLAB crushed with a segmentation error, so the average carea under the curve is computed only for the two first object classes.

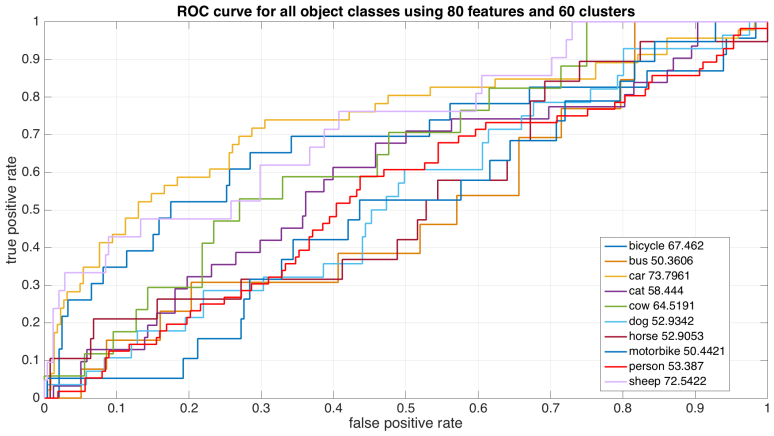

Figure 1: ROC curve for 80 Features Descriptors and 60 Clusters

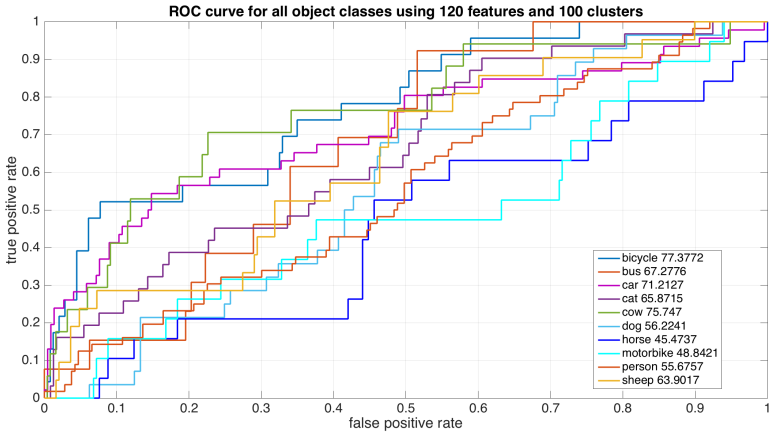

Figure 2: ROC curve for 120 Features Descriptors and 100 Clusters

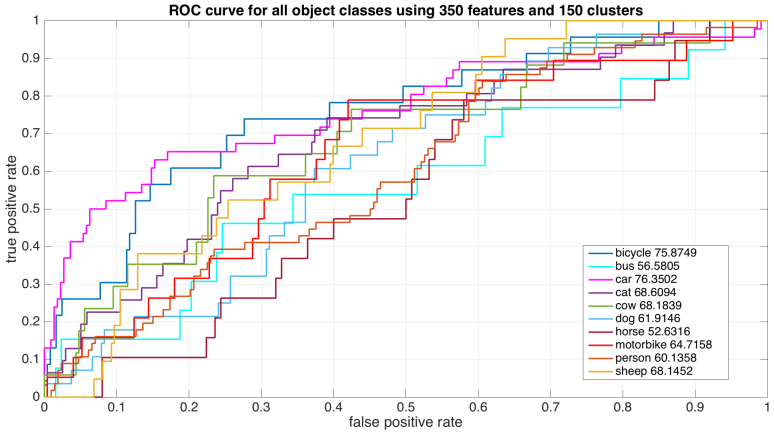

Figure 3: ROC curve for 350 Features Descriptors and 150 Clusters

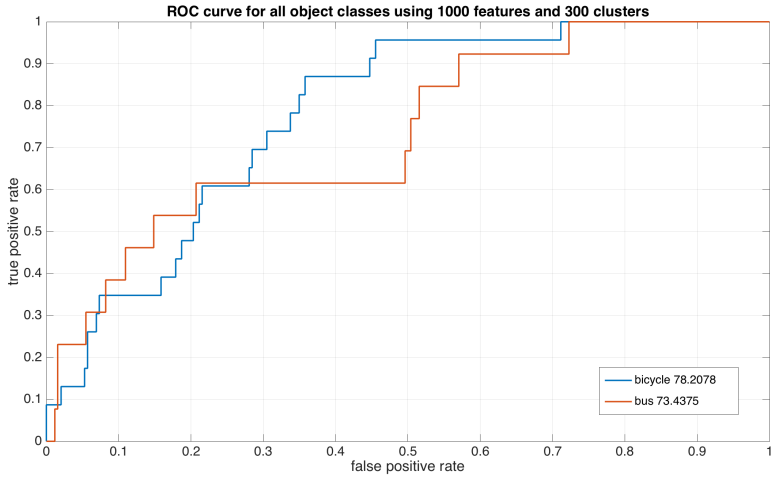

Figure 4: ROC curve for 1000 Features Descriptors and 300 Clusters

Table 2: Average AUC for Each of the Classifiers

	Average AUC
classifier 1	59.68 %
classifier 2	62.76 %
classifier 3	65.31 %
classifier 4	75.82 %

4 Discussion

From the Figures in the Results section is not very apperent, but the average area under the curve shown in the table for each of the classifiers shows that the accuracy increases with the number of feature detetctors and the number of clusters, while the time cost also increases. Still the results are not perfect. One way of improving the results would be to make the Laplacian of Gaussian rotation and affine invarient, since the code that is used for detecting the keypoints it is only scale invariant. This should improve the results by some percent, but also it should be taken into account the

time elapsed when extracting the features. Therefore an optimal number of features to be extracted and number of clusters should be chosen, and since the feature extraction will be rotational and affine invariant, fewer fetures will give better results.

5 Conclusion

Object classification was performed using a linear SVM classifier and the method of bag of words using local descriptors. Although the method proposed in this paper is time costy, the results show that the accuracy of the classifier increases with the number of feature descriptor extracted and the number of clusters when constructing the dictionary.

References

[1] Lindeberg, T. (1998). Feature Detection with Automatic Scale Selection. *IEEE Transactions Pattern Analysis Machine Intelligence*, 30.

[2] Nowak, E., Jurie, F., & Triggs, B. (2006). Sampling Strategies for Bag-of-Features Image Classification. *Computer Vision – ECCV 2006 Lecture Notes in Computer Science*, 490-503. doi:10.1007/11744085_38

[3] Mikolajczyk, K., & Schmid, C. (2005). A performance evaluation of local descriptors. it IEEE Computer Society Conference on Computer Vision and Pattern Recognition, doi:10.1109/cvpr.2003.1211478

[4] Hassaballah, M., Abdelmgeid, A. A., & Alshazly, H. A. (2016). Image Features Detection, Description and Matching. *Image Feature Detectors and Descriptors Studies in Computational Intelligence*, 11-45. doi:10.1007/978-3-319-28854-3_2