Bag-of-Words based Image Classification

Rémi de Zoeten and Anouk Visser

December 27, 2013

1   Introduction

In this report we will discuss the implementation and results of a system for image classi-fication. The system can tell if there is an object of one of four given classes (motorbikes, cars, faces and airplanes) in an image. The classification system is based on a Bag-of-Words approach, meaning that the Support Vector Machine that performs the classification uses histograms of words for training and classification of an image. The words are obtained by extracting descriptors from a set of training images and cluster them using the *kmeans* algorithm.

2   Implementation

We have extracted key point SIFT features and dense SIFT features. In addition to this we have also trained and evaluated our system using RGBSIFT, rgbSIFT and opponentSIFT, these are just key point- or dense SIFT features applies to all three of the color channels of the specific colorspace, resulting in descriptors of dimension 128×3. For dense SIFT sampling we used a bin size of 10 pixels and a step size of 5.

To build a visual vocabulary, we use the extracted descriptors from a number of training images (of all classes) and use this as the data for the kmeans clustering algorithm. The number of clusters represents the number of words in the vocabulary. Because of our limited computational resources the largest vocabulary we have built is a vocabulary of 100 words based on 5 training images.

We will then use the words to quantize training and test images. We do this by extracting descriptors from the image and find the descriptor from the vocabulary with the smallest euclidean distance with respect to the descriptor from the image. We will keep track of which words (i.e. descriptors from the vocabulary) were selected and how many times this happened in a histogram. When every descriptor has been 'matched' to a word from the vocabulary, we normalize the resulting histogram by dividing every bin by the total number of descriptors extracted from the image. We perform this quantization on both the images for training and testing.

The histograms of the training images are then used to train a Support Vector Machine. We train one SVM for every class, for the 'motorbikes' we will use the histograms of training images of a motorbike as positive examples and histograms of the other images as negative examples. After we have trained the classifiers we can use these to make predictions about the test histograms.

As an evaluation measure we have used the Mean Average Precision. We provide every classifier with all the test images and rank the results based on the probability that the test image is off the class the classifier is trained on. For every classifier we will thus have a ranked list of test images, if everything works well the top of the list should be dominated by images of the class the classifier can predict. In the next section we will show the results of our system and provide an analysis of these results.

3 Evaluation

3.1 Size of the training set

The first experiment we will discuss measures the performance of the system for all different methods of SIFT (dense and key point combined with all colorspaces results in eight methods). We have set out the Mean Average Precision against the number of images that were used to train the SVM. The results of this experiment can be found in figure 1. The first thing that stands out is the rather large difference between the key point SIFT features and the dense SIFT features. It appears that dense SIFT features perform much better than key point SIFT features. Performance of all methods increases rapidly up until using approximately 40 training images. However, after this point there is only a slight increase in performance when increasing the training size. Only key point SIFT applied to the rob colorspace shows a sudden peak in performance, we are not sure whether this is actually happening or that we can consider this as noise in the data.
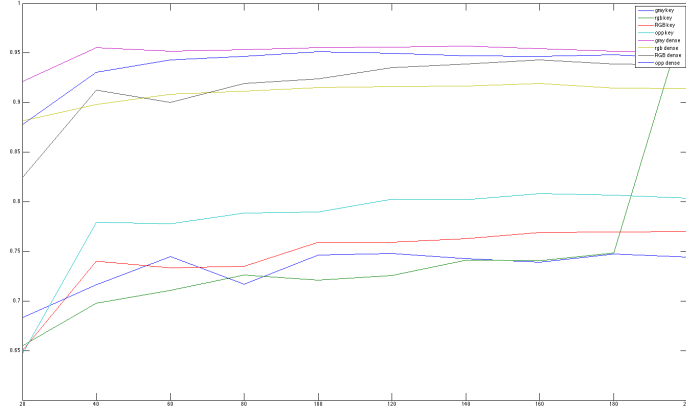
Figure 1: Results of the system for different sizes of the set of image histograms used to train the SVM. The SVM was trained using an RBF kernel. For this experiment we have used a vocabulary size of 50 that was made by clustering the descriptors of 10 images from each training class.

3.2   Size of the vocabulary

Similar to the experiment described in section 3.1 we have performed an experiment in which we measured performance for all different methods under varying vocabulary sizes, the results of which are shown in figure 2. Again we find that the dense SIFT features perform better than the key point SIFT features, with the exception of the key point SIFT applied to the rgb color space. The rgb color space shows excellent performance for all different vocabulary sizes. We are unsure why this is the case, but after careful inspection of the classifiers trained on these features we have concluded that it really performs this well and that it is not just noise. We observe that the majority methods benefit from a larger vocabulary size.
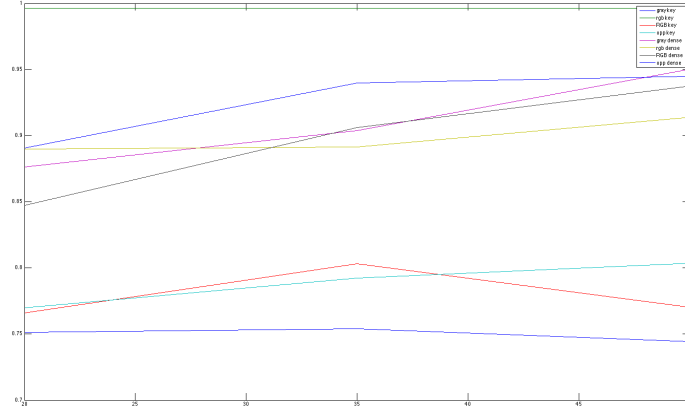
Figure 2: Results of the system for different sizes of the vocabulary. For this experiment we the descriptors of 10 images from each training class to build the vocabulary. We used 200 image histograms from each class to train the SVM. The SVM was trained using an RBF kernel.

3.3 Kernel functions for the SVM

4 Conclusion

In this report we have discussed our implementation of a system for image classification based on a Bag-of-Words approach. We have shown its performance for different sizes of the training data, different vocabulary sizes, different SIFT features for different colorspaces and different kernels used to train the SVM. Unfortunately we did not have the computing power to perform any larger tests (for instance larger vocabularies based on larger training sets). However, we found that the results using smaller vocabulary sizes were above expectations.