University of Amsterdam

Computer Vision 1

Final Assignment

Automated Image Classification

Elise Koster, 5982448
Sharon Gieske, 6167667
David van Erkelens, 10264019

Master's Programme Artificial Intelligence
Graduate School of Informatics
University of Amsterdam

October 27, 2014

1. Introduction

Object classification is a fundamental part of Computer Vision and can be used to automate processes and provide environmental reasoning for robotics. Consequently, a high accuracy in object classification systems is invaluable for industry and science alike.
This paper outlines the results of a project analyzing different approaches to image classification using techniques from Machine Learning and Computer Vision. Multiple techniques have been implemented and tested, yielding different results.
The data section will describe the images used for training and testing, the implementation section will introduce the techniques used and the results section will describe the difference in performance for each set of techniques. The conclusion will report on the optimal combination found and the section future work will comment on possible improvements.

2. Data

The training data consists of 2000 .jpg-images in four classes (500 per class): airplanes, cars, faces and motorbikes. The group of training images is split into a vocabulary-training set and a classification training set.The test data consists of 200 .jpg-images in the same four classes, all of whom are used for testing.

3. Implementation

Instead of training classifiers on a large set of pixels, the bag-of-words approach is used. This approach first extracts features from images and subsequently uses them to build a vocabulary of visual 'words'. Each image can then be described as a set of these words, which makes training a classifier easier and faster than a pixel-by-pixel approach.

3.1. SIFT

To be able to build a bag-of-words, features need to be extracted from each training image. This is done using Scale Invariant Feature Transform (SIFT), an algorithm that detects points of interest in an image and produces descriptors of these features. Two types of SIFT are used: key-point (produces descriptors of points of interest) and dense-sampling (every n pixels a descriptor is produced). A multitude of color spaces is used: gray-scale, RGB (regular .jpg-image with three channels), normalized RGB (rgb) and opponent, where R, G, B respectively

are the pixel values per color channel, and

For rgb:

$$r = \frac{R}{R+G+B}$$
$$g = \frac{G}{R+G+B}$$
$$b = \frac{B}{R+G+B}$$

For opponent:

$$O_1 = \frac{R-G}{\sqrt{2}}$$
$$O_2 = \frac{R+G-2B}{\sqrt{6}}$$
$$O_3 = \frac{R+G+B}{\sqrt{3}}$$

Each different color space defines different intensities for each pixel in a color channel, and thus causes SIFT to return different descriptors.

3.2. K-means

Performing SIFT on the first set of training images results in a set of descriptors for each image, which are clustered into visual words using K-means. The K-means algorithm works by calculating the Euclidian distance between each data point and a cluster-center (the mean), and iteratively re-calculating the means and re-assigning the data-points until convergence. The resulting clusters form the visual vocabulary used for describing and classifying images later on.

3.3. SVM classification

Once the visual vocabulary has been built, features are extracted from a new set of training images. These features are grouped into words according to the visual vocabulary, and for each image a histogram of visual word frequencies is computed. These histograms are used as input to train four Support Vector Machines (SVMs) (one per class), using different kernel-functions. SVMs are non-parametric classifiers, which work by maximizing the margin between the decision boundary and two classes of data. After training, all test images are classified according to the SVM-models built using the training images.

4. Results

Results of:

- key points vs dense
- vocabulary size(400,800,1600,2000,4000)
- SIFT color space (gray-scale, RGB, rgb, opponent)
- amount of training samples used (vocab)
- amount of training samples used (svm)
- kernel choice for sum

SIFT type	Samples:vocabulary	Samples:SVM	Clusters	MAP
Key, grayscale	50	50	400	0.7041
Key, grayscale	50	100	400	0.7362
Key, grayscale	50	150	400	0.7322
Key, grayscale	50	250	400	0.7518
Key, grayscale	50	300	400	0.7548

Table 1. Mean Average Precision for different training sizes, 400 clusters

Airplanes	Cars	Faces	Motorbikes
0.7219	0.7232	0.5326	0.8386
0.7937	0.7525	0.5547	0.8441
0.7419	0.7914	0.5450	0.8503
0.8226	0.7722	0.5415	0.8710
0.8311	0.7612	0.5521	0.8748

Table 2. Average precision per class for table 1

SIFT type	Samples:vocabulary	Samples:SVM	Clusters	MAP
Key, grayscale	50	50	800	0.5911
Key, grayscale	50	100	800	0.6161
Key, grayscale	50	150	800	0.6235

Table 3. Mean Average Precision for different training sizes, 800 clusters

Airplanes	Cars	Faces	Motorbikes
0.5978	0.6869	0.4507	0.6292
0.5432	0.7424	0.3978	0.7811
0.6434	0.5742	0.4207	0.8557

Table 4. Average precision per class for table 3

5. Conclusion

6. Future work

In the current implementation, SIFT for color spaces uses an average over color channel histograms. A possible improvement (though less computationally tractable) would train an SVM per color channel and build a committee of these SVMs to classify the image.