COMPUTER VISION 1 - FINAL PROJECT

SHARON GIESKE, ELISE KOSTER, DAVID VAN ERKELENS

Introduction

Object classification is a fundamental part of Computer Vision and can be used to automate processes from factory lines to traffic control. Consequently, a high accuracy in object classification systems is invaluable to industry.
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, the results section will describe the difference in performance for each set of techniques, and the conclusion will report on the optimal combination found.

1. Data

The training data consists of 2000 .jpg-images in four classes (500 per class): airplanes, cars, faces and motorbikes. The test data consists of 200 .jpg-images in the same four classes.

2. 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.
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). SIFT is used against a multitude of color spaces: gray-scale, RGB (regular .jpg-image with three channels), normalized rgb (where $r = \frac{R}{R+G+B}$, $g = \frac{G}{R+G+B}$ and $b = \frac{B}{R+G+B}$) and opponent (where $O_1 = \frac{R-G}{\sqrt{2}}$, $O_2 = \frac{R+G-2B}{\sqrt{6}}$ and

$O_3 = \frac{R+G+B}{\sqrt{3}}$) First, features are extracted from a set of training images using key-point and dense-sampling SIFT, for gray-scale, RGB, normalized RGB (rgb) and opponent color spaces.
This results in a set of descriptors for each training image, which are clustered into visual words using K-means. The resulting clusters form a visual vocabulary.
Then, 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 SVM-classifiers (one per class), using different kernel-functions. After training, all test images are classified according to the SVM-models built using the training images.

3. 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	color space	accuracy
x	y	z
a	b	c

4. Conclusion