Visual Recognition, Fall 2011

Implementation assignment

Due: Friday, Sept 16

Overview

For this assignment, the goal is to implement a bag-of-words image recognition system, and then evaluate it on natural images containing four object categories. Feel free to use existing software for any of the steps below (when available); cite any components you have borrowed in your writeup.

Dataset

Download the image data from the course webpage. The classes are airplanes, motorcycles, faces, and cars. The file has train and test subdirectories per category, with 50 test images per class and up to 500 training images per class.

System components

Below are the main steps you will need to include. Aside from following this basic pipeline, the details for the implementation are up to you.

1. **Detection**: Determine where in the image to extract local feature descriptors. You could use a single interest operator (e.g., DoG, MSER, Harris), or a combination of interest operators, or else forgo interest point selection and densely sample feature points.

2. **Description**: Choose a local image descriptor to extract at each of the selected points (e.g., SIFT, SURF, pixel intensities in the patch).

3. **Visual vocabulary**: Sample some descriptors from the training images, and apply k-means (or some other approach) to them to quantize the feature space into “visual words”.

4. **Bag of words formation**: Map the set of features from an image into its bag-of-words histogram, using the visual vocabulary constructed above. Do this for all training images and all test images.

5. **Training**: Build a multi-class classifier of your choice (e.g., SVM, Naïve Bayes, k-nearest neighbors,…). Use the training partition provided (as given by the subdirectories).

6. **Testing**: Apply the learned classifier to categorize each test image into one of the four categories.

Evaluation and analysis

Write a report that includes the following:

- A description in English (not code) of precisely how your approach was implemented, listing the steps and details of any choices you made.
- A visualization of a couple visual words using examples from the data.
- Report the classification accuracy – the percentage of images in the test set that are correctly classified by your program.
- Compute and display a confusion matrix, and explain the result.
- Discuss the overall performance of your system.
- Test two variants of your system, and design comparative experiments to show how they differ. Carefully *explain* the impact on accuracy. These two variants could be based on adjusting choices you made in 1:6 above, or some other feature that you optionally add to the system (e.g., using segmentation, context, feature selection, spatial pyramid match kernels, vocabulary-trees). Illustrating the impact on accuracy could go beyond the overall accuracy number, such as impact on particular classes, types of examples, etc. *Note: please avoid simply reporting the impact of tuning parameters, unless you have a good conceptual explanation for the impact*.

Grades will be determined based on implementation completeness as well as the insights demonstrated by the report discussion. Use figures or images to make your point clear when appropriate.

What to submit

Attach your report (as a single pdf, please) to an email to Kristen, with [395T] in the subject line.

References

See the course webpage http://www.cs.utexas.edu/~grauman/courses/fall2011/schedule.html for readings and many links to useful code (for interest point detection, SIFT, SVM libraries, clustering, etc.) These papers are especially relevant:

- Visual Categorization with Bags of Keypoints, Dance et al., ECCV International Workshop on Statistical Learning in Computer Vision, 2004.
- Sampling Strategies for Bag-of-Features Image Classification. Nowak et al. ECCV 2006.
- Video Google: A Text Retrieval Approach to Object Matching in Videos, Sivic and Zisserman, ICCV 2003.
- Object Recognition from Local Scale-Invariant Features, Lowe, ICCV 1999.

Acknowledgements

The dataset is provided courtesy of the Caltech Vision Group, and collated by Svetlana Lazebnik.