Laboratory Experience 2 - Optional

Task This experience focuses on a visual recognition algorithm called Naive-Bayes Nearest-Neighbor [1]. You are asked to implement the algorithm and test it on a scene recognition dataset.

A short explanation of the algorithm. Given a query image $I = \{d_1, \ldots, d_n\}$ (where $d_i \in \mathbb{R}^D$ is a local image descriptor) and a set of classes $\mathbf{C} = \{C_1, \ldots, C_n\}$, the ML estimate of the class of image I is:

$$\hat{C} = \arg\max_{C \in \mathbf{C}} p(I|C) = p(d_1, \ldots, d_n|C). \tag{1}$$

This also corresponds to the MAP estimate $\arg\max_{C \in \mathbf{C}} p(C|I)$, whenever the class priors $p(C)$ are uniform. Taking the negative logarithm of this quantity and using the Naive-Bayes assumption (that the local descriptors are conditionally independent, given the class C), we obtain:

$$\hat{C} = \arg\min_{C \in \mathbf{C}} -\log p(d_1, \ldots, d_n|C) \tag{2}$$

$$= \arg\min_{C \in \mathbf{C}} -\log \prod_{i=1}^{n} p(d_i|C) \tag{3}$$

$$= \arg\min_{C \in \mathbf{C}} -\sum_{i=1}^{n} \log p(d_i|C). \tag{4}$$

We can estimate $p(d_i|C)$, using a kernel density estimator:

$$\hat{p}(d_i|C) = \frac{1}{Lh^D} \sum_{l=1}^{L} K\left(\frac{d_i - d_{lC}}{h}\right), \tag{5}$$

where d_{jC} is the j-th local descriptor from class C, L is the total number of local descriptors in C, $K(x) = (2\pi)^{-\frac{D}{2}} \exp\left(-\|x\|^2\right)$ and h is the bandwidth parameter.
This quantity is difficult to compute, because the number of local descriptors in a class C is huge. Nonetheless it can reliably be approximated [1] by using only the single Nearest

Neighbor d_{NN_iC} of d_i in class C, to obtain the final classification rule:

$$\hat{C} = \arg\min_{C \in \mathbf{C}} - \sum_{i=1}^{n} \log \hat{p}(d_i|C) \tag{6}$$

$$= \arg\min_{C \in \mathbf{C}} - \sum_{i=1}^{n} \log \left(\frac{1}{Lh^D} \sum_{l=1}^{L} K \left(\frac{d_i - d_{lC}}{h} \right) \right) \tag{7}$$

$$\approx \arg\min_{C \in \mathbf{C}} - \sum_{i=1}^{n} \log K \left(\frac{d_i - d_{NN_iC}}{h} \right) \tag{8}$$

$$= \arg\min_{C \in \mathbf{C}} \sum_{i=1}^{n} \|d_i - d_{NN_iC}\|^2 \tag{9}$$

The resulting classification algorithm is extremely simple, requires no training and it can achieve classification performances comparable to the more complicate bag of visual words models.

Experiments

1. Implement the NBNN algorithm:
   - to efficiently compute the Nearest-Neighbor d_{NN_iC} in class C of a descriptor d_i, it's necessary to make use of an approximate NN search algorithm
   - we suggest to make use of FLANN, a widely used open source library for approximate Nearest Neighbor, with a Matlab interface. You can download it from: http://mloss.org/software/view/143/
   - remember that this algorithm requires using the local SIFT descriptors (i.e. /path/to/15Scenes/features/SIFT(...).mat), rather than the PHOW features
2. Perform a scene recognition experiment on the 15 Scenes dataset [2], using:
   - the same experimental protocol of the first mandatory experience
   - the features already computed for the first mandatory experience
   - $\alpha = \{0, 1\}$ (the coefficient of the spatial coordinates)
3. Optionally, experiment by decreasing the spacing between the SIFT local descriptors (e.g. to 6, or 4 pixels)
4. Optionally, experiment with a configuration of your choice on the ISR dataset [3], using the features and the experimental protocol introduced in the first mandatory experience

Run each experiment twice, with different training/testing splits and report the multiclass accuracies (the mean class recognition rate), as mean $\pm$ std.
For the best configuration report also:

- the confusion matrix
- recognition rate per class

References

[1] O. Boiman, E. Shechtman, and M. Irani. In defense of nearest-neighbor based image classification. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE, 2008.

[2] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, volume 2, pages 2169–2178. IEEE, 2006.

[3] A. Quattoni and A. Torralba. Recognizing indoor scenes. In *In Proc. Computer Vision and Pattern Recognition*. IEEE, 2009.