Learning Tree-structured Descriptor Quantizers for Image Categorization

Josip Krapac[1]
http://lear.inrialpes.fr/~krapac
Jakob Verbeek[1]
http://lear.inrialpes.fr/~verbeek
Frédéric Jurie[2]
http://users.info.unicaen.fr/~jurie

[1] LEAR Team, INRIA Grenoble, Montbonnot St-Martin, France
[2] Université de Caen Basse-Normandie, Caen, France

Overview Current state-of-the-art image categorization systems rely on bag-of-words representations that model image content as a histogram of quantization indices that code local image appearance. Usually the quantizer is constructed in unsupervised way, *e.g.* using k-means quantization or Gaussian mixture models. Although sometimes it might be advantageous to have a generic image representation that can be used to address different tasks, the fact that the representation is not optimized for a specific task means that it is suboptimal in the sense that it less discriminatively and/or less succinctly captures the relevant image content.
In this work we construct the quantizer in a supervised way, using tree-structured quantizers that allow computationally efficient assignment of patch descriptors to quantization indices. We optimize the quantizer for image classification, differently from [4] which optimizes tree-structured quantizers for patch classification, where the class labels for patches are inherited from images.
Our quantizers are fast and efficient. The structure allows fast image representation creation, and the use of linear models to obtain the score for an image ensures fast classification given an image representation. Very compact representations perform already excellent, comparable to the performance obtained using k-means vocabularies of an order of magnitude larger.

Quantizer construction We quantize the feature space using binary decision trees. Each non-leaf node n has an associated split criterion $f(x;\theta_n,\tau_n)$ whose sign determines to which child a given feature vector x will proceed and (θ_n,τ_n) denote the split parameters of node n. Each node of the tree corresponds to a part of feature space: the root is associated with the complete feature space, child nodes being associated with a subset of the space associated with the parent. An image is represented by an l_1 normalized histogram that codes in dimension d the fraction of the image regions associated with the d-th leaf.
We start a tree construction with a trivial tree with just the root-node and expand the existing tree by splitting the leaf-nodes. Splitting a leaf-node refines the partitioning of the feature space, therefore expanding the current image representation. At each tree expansion step, we sample T candidate splits, by uniformly selecting one of the existing leafs, and by sampling the split criterion parameters (θ,τ). We used simple split criterion that compares a value of feature at a selected dimension with a selected threshold, but more general split functions are possible. Each node split corresponds to candidate expansion of the current tree and induces a new image representation: the histogram bin corresponding to the parent node being split is replaced by two new histogram bins corresponding to the two new child nodes. We learn a classifier on a train set of images, using this enlarged image representation. We employ a greedy procedure by accepting the split which induces the best classification performance on a validation set of images. We iteratively continue splitting the leaf-nodes until we reach the specified number of leaves.
We evaluate performance in terms of average precision (AP), so we aim to optimize ranking performance rather than classification performance. Therefore we learn a linear score function for ordinal regression where the goal is to ensure that for each pair of a positive and a negative image the score of the positive one is larger than the score of the negative one by some margin. We obtain the score function parameters by solving optimization problem [2] using the cutting-plane method.
We have found that using an ensemble of several trees gives significant improvements, as has also been observed before for randomized decision trees [1]. Here we learn each tree independently, and then concatenate the histogram representations obtained using each tree to obtain our final image representation. Combining K trees of L leafs each we obtain a final representation of size $K \times L$. In addition to improving the results, using ensembles also significantly reduces the variance of performance.

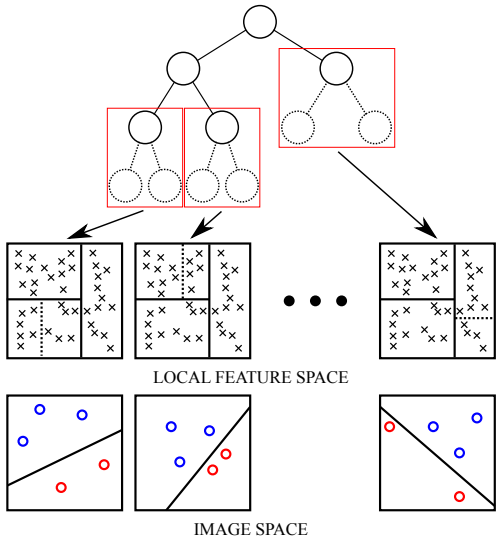

Figure 1: The split of a node in the quantization tree refines the quantization of the feature space, and replaces one dimension in the image representation space with two new ones, improving the classification rate.

Experimental evaluation We compare to tree-structured quantizers of [4], k-means ensebles and larger k-means vocabularies on Graz-02 and 15-scenes datasets. On Graz-02 we outperform the baselines for all numbers of quantizers in ensemble, and all number of quantization cells, except for one class where four times bigger k-means vocabularies are better. On 15-scenes dataset we improve less over the patch-based trees, probably because here the complete image contains useful information, so the patch-based trees do not wastefully model irrelevant background features.

AP×100	K	L	bicycle	car	person
k-means	10	10	82.0 ± 1.2	66.4 ± 2.1	68.4 ± 2.3
patch-based tree	10	10	78.8 ± 3.1	75.9 ± 2.4	68.6 ± 2.9
image-based tree (ours)	10	10	85.8 ± 1.8	81.3 ± 0.4	78.0 ± 0.7
k-means	10	100	86.4 ± 1.3	80.1 ± 2.5	76.5 ± 0.5
patch-based tree	10	100	86.8 ± 1.0	82.7 ± 2.0	77.1 ± 2.0
image-based tree (ours)	10	100	**91.2 ± 0.6**	**87.5 ± 0.8**	85.3 ± 0.9
k-means	1	1000	88.3 ± 1.9	81.1 ± 0.8	83.1 ± 0.7
k-means	1	2000	90.0 ± 0.3	83.1 ± 0.6	85.6 ± 0.4
k-means	1	4000	90.7 ± 0.3	84.8 ± 1.2	**87.3 ± 0.4**

Table 1: Summary of the results on Graz-02 dataset.

	K	L	accuracy	mAP
k-means	10	10	59.7 ± 0.4	57.3 ± 0.6
patch-based tree	10	10	79.0 ± 0.9	70.0 ± 0.7
image-based tree	10	10	80.0 ± 0.9	78.9 ± 0.5
k-means	10	100	73.7 ± 0.8	79.2 ± 0.7
patch-based tree	10	100	**83.9 ± 0.6**	84.2 ± 0.8
image-based tree	10	100	**83.6 ± 0.6**	**85.6 ± 0.5**
k-means	1	1000	80.5 ± 0.7	84.0 ± 0.8
Lian *et al*. [3]	105	50	78.1 ± 0.7	–

Table 2: Summary of the results on 15-scenes dataset.

[1] L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.
[2] T. Joachims. A support vector method for multivariate performance measures. In *ICML*, 2005.
[3] X. Lian, Z. Li, B. Lu, and L. Zhang. Max-margin dictionary learning for multiclass image categorization. In *ECCV*, 2010.
[4] F. Moosmann, E. Nowak, and F. Jurie. Randomized clustering forests for image classification. *PAMI*, 30(9):1632–1646, 2008.