Learning Directional Local Pairwise Bases with Sparse Coding

Nobuyuki Morioka
nmorioka@cse.unsw.edu.au
The University of New South Wales & NICTA, Sydney, Australia

Shin'ichi Satoh
satoh@nii.ac.jp
National Institute of Informatics, Tokyo, Japan

Recently, sparse coding has been receiving much attention in object and scene recognition tasks because of its superiority in learning an effective codebook over k-means clustering [5]. However, empirically, such codebook requires a relatively large number of visual words, essentially bases, to achieve high recognition accuracy. Therefore, due to the combinatorial explosion of visual words, it is not practical to use sparse coding to represent higher-order spatial features which are equally important in capturing distinct properties of scenes and objects.

Contrasted with many previous techniques that exploit higher-order spatial features [1], Local Pairwise Codebook (LPC) is a simple and effective method to learn a compact set of clusters representing pairs of spatially close descriptors with k-means [4]. Based on LPC, we propose Directional Local Pairwise Bases (DLPB) that applies sparse coding to learn a compact set of bases capturing correlation between these descriptors, so to avoid the combinatorial explosion. Furthermore, such bases are learned for each quantized direction thereby explicitly adding directional information to the representation.

Given an image, we densely sample feature points and each feature f_i is encoded as $(x_i, y_i, \mathbf{d}_i)$ where x_i and y_i denote the feature location and $\mathbf{d}_i$ is the feature descriptor. This is followed by pairing up the features that are within δ pixels away from each other. For each pair of spatially close descriptors, we extend the joint descriptor representation of LPC by assigning one of the discretized directional relationships illustrated in Figure 1. We set the partition of each directional kernel to be roughly equal to each other.

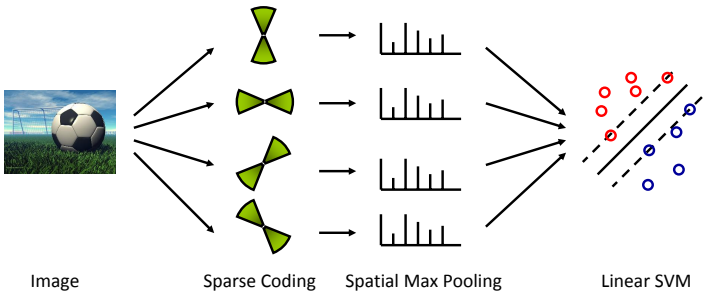

Figure 1: Directional Local Pairwise Bases. Given a feature descriptor in an image, four directional relationship kernels are applied to form local pairwise descriptors with nearby descriptors. For each direction, sparse coding is applied and statistics are computed based on spatial max pooling. The statistics for all four directions are then concatenated and fed into a linear SVM classifier

Once r_{ij} is determined, each pair of spatially close descriptors is represented as:

$$f_{(i,j)} = \left(\frac{(x_i + x_j)}{2}, \frac{(y_i + y_j)}{2}, r_{ij}, \mathbf{p}_{ij} \right) \tag{1}$$

where $\mathbf{p}_{ij}$ denotes vector concatenation of two feature descriptors $[\mathbf{d}_i \mathbf{d}_j]$. The order of $\mathbf{d}_i$ and $\mathbf{d}_j$ is based on the spatial locations of the features and their assigned directional relationship. We simply compare y-coordinates when the directional relationship is vertical and x-coordinates for the other three directions.

Then, we learn a specific dictionary for each directional relationship r with sparse coding. If we denote a set of possible directional relationships as $R = \{vt, hz, d1, d2\}$, then the learning problem is stated as follows:

$$\min_{D,\alpha} \sum_{r \in R} \sum_{i}^{N^r} \|\mathbf{p}_i^r - D^r \alpha_i^r\|_2^2 + \lambda \|\alpha_i^r\|_1 \tag{2}$$
$$\text{subject to } \|D_j^r\|^2 \leq c, \forall j = 1, \ldots, k.$$

where the sparse representation of a pairwise feature descriptor $\mathbf{p}_i^r$ with a directional relationship r is α_i^r and is inferred with a direction specific dictionary D^r. N^r is the number of pairwise feature descriptors sampled to learn D^r. Since each set $\{D^r, \alpha^r\}$ can be learned and inferred independently from each other, the computational cost is significantly reduced if a large dictionary D is required. To learn each dictionary, we have adopted an efficient sparse coding algorithm proposed by Lee et. al. [3].

Once the dictionaries are learned, we represent each image with spatial max pooling. This is done by first inferring a sparse code vector α_i^r for each local pairwise feature descriptor $\mathbf{p}_i^r$ formed in an image. Then, we obtain a set of local statistics over multiple scales. At every level of scale l, the image is partitioned into $M_l \times M_l$ disjoint regions where M_l is defined as 2^{l-1}. For each region m, its statistics are represented as a vector $\mathbf{h}_{l,m}^r$ where each dimension j stores the max value of each basis $\mathbf{D}_j^r$ spatially pooling all pairwise descriptors with a directional relationship r within the region m as stated below:

$$\mathbf{h}_{l,m}^r(j) = \max\{|\alpha_1^r(j)|, |\alpha_2^r(j)|, \ldots, |\alpha_{N_m}^r(j)|\} \tag{3}$$

where N_m denotes the number of descriptors inside the region m. Once the local statistics are computed for all regions at scale l, the vectors are concatenated and are ℓ_2 normalized as:

$$\mathbf{h}_l^r = \frac{[\mathbf{h}_{l,1}^r \mathbf{h}_{l,2}^r \ldots \mathbf{h}_{l,m}^r]}{\|[\mathbf{h}_{l,1}^r \mathbf{h}_{l,2}^r \ldots \mathbf{h}_{l,m}^r]\|_2} \tag{4}$$

To obtain the final representation of the image denoted as $\mathbf{h}$, we further concatenate all $\mathbf{h}_l^r$ over multiple levels of scale ($l \in \{1 \ldots L\}$) and multiple dictionaries ($r \in R$). Finally, we train a model with a linear SVM and use the learned model for classifying test images.

We have evaluated DLPB against SPM [2], ScSPM [5] and DLPC, a variant of DLPB that uses k-means clustering instead of sparse coding to learn direction specific codebooks. A summary of experimental results on five challenging datasets is given in Table 1. It is clear that DLPB has consistently outperformed the three baseline methods.

Methods	C101	C256	MSRCv2	Pascal	15SC
SPM	65.38	29.14	84.4	46.7	81.76
ScSPM	71.98	35.19	87.0	51.9	84.32
DLPC	71.60	36.74	86.4	52.6	83.80
DLPB	**77.33**	**40.81**	**88.5**	**54.0**	**85.22**

Table 1: A summary of experimental results on Caltech-101 (C101), Caltech-256 (C256), MSRCv2, Pascal VOC 2007 (Pascal) and 15 Scenes (15SC)

[1] S. Lazebnik, C. Schmid, and J. Ponce. A Maximum Entropy Framework for Part-Based Texture and Object Recognition. In *ICCV*, 2005.

[2] S. Lazebnik, C. Schmid, and J. Ponce. Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories. In *CVPR*, 2006.

[3] H. Lee, A. Battle, R. Raina, and A.Y. Ng. Efficient sparse coding algorithms. In *NIPS*, 2006.

[4] N. Morioka and S. Satoh. Building Compact Local Pairwise Codebook with Joint Feature Space Clustering. In *ECCV*, 2010.

[5] J. Yang, K. Yu, Y. Gong, and T.S. Huang. Linear spatial pyramid matching using sparse coding for image classification. In *CVPR*, 2009.