Multipath Sparse Coding Using Hierarchical Matching Pursuit

Yufeng Jiang

1. Hierarchial Matching Pursuit

In their hierarchical matching pursuit, MI-KSVD is used to learn codebooks at three layers, where the data matrix Y in the first layer consists of raw patches sampled from images, and Y in the second and third layers are sparse codes pooled from the lower layers. With the learn codebooks D, hierarchical matching pursuit builds a feature hierarchy, layer by layer, using batch orthogonal matching pursuit for computing sparse codes, spatial pooling for aggregating sparse codes, and contrast normalization for normalizing feature vectors, as shown in Fig. 1 used in [2] **First Layer:** The goal of the layer in HMP is to extract sparse codes for small patches (*e.g.* 5×5) and generate pooled codes for mid-level patches (*e.g.* 16×16). Orthogonal matching pursuit is used to computed the sparse codes x of small pathces (*e.g.* 5×5 pixels) The features of each spatial cell C are the max pooled sparse codes, which are simple the component-wise maxima over all sparse codes within a cell [2]:

$$F(C) = \max_{j \in C}[max(x_{j1}, 0), \ldots, max(x_{jM}, 0), \ldots, \quad (1)$$
$$max(-x_{j1}, 0), \ldots, max(-x_{jM}, 0)]$$

Here, j ranges over all entries in the cell, and x_{jm} is the m-th component of the sparse code vector x_j of entry j. Authors split the positive and negative components of the sparse codes into separae features to allow higher layers weight positive and negative reponses differently. The feature F_P describing an image patch P is the concatenation of aggregated sparse codes in each spatial cell [2]

$$F_P = [F(C_1^P), \ldots, F(C_s^P), \ldots, F(C_S^P)] \quad (2)$$

where $C_s^P \subseteq P$ is a spatial cell geneated by spatial partitions, and S is the total number of spatial cells. They additionally normalize the feature vectors F_P by L_2 norm $\sqrt{||F_P||^2 + \varepsilon}$, where ε is a small positive number. Since the magitude of sparse codes varies over a wide range due to local variations in illumination and occlusion, this operation makes the appearance features robust to such variations.

Second Layer: The goal of the second layer in HMP is to gather and code mid-level sparse codes and generate pooled codes for large patches (*e.g.* 36×36). HMP applies batch OMP and spatial max pooling to features F_P generated in the first layer.

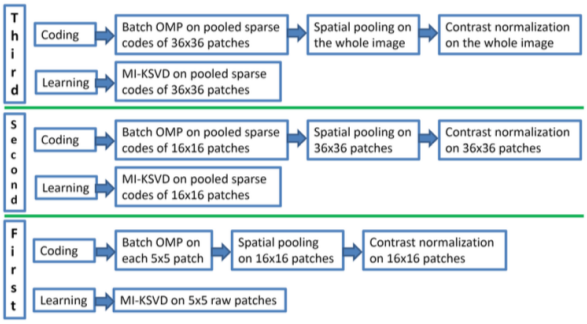

Figure 1. A three-layer architecture of Hierarchical Matching Pursuit.

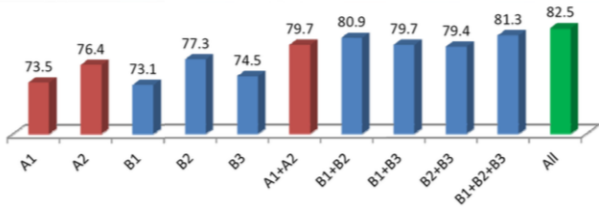

Figure 2. Test accuracy of single-path and multipath HMP. A1,A2,B1,B2 and B3 denotes the HMP networks of different architectures. A1+A2 indicates the combination of A1 and A2. "All" means the combination of all five paths: A1,A2,B1,B2 and B3.

Third Layer: The goal of the third layer in HMP is to generate pooled sparse codes for the whole image. Similar to the second layer, the codebook for this level is learned by sampling these pooled sparse codes in the second layer.

2. Experiments

Authors evaluate the proposed M-HMP models on five standard vision datasets on object, scene and fine-grained recognition, extensively comparing to state-of-the-art algorithms using designed and learned features. All images are resized to 300 pixels on the longest side.

2.1. Object Recognition

They investigate the behavior of M-HMP for object category recognition on Caltech-101. The results of M-HMP are shown in Fig. 2 cited in [2]. As can been seen, the one-

SIFT+T [5]	67.7	HSC [11]	74.0
Local NBNN [8]	71.9	Asklocals [3]	77.1
LC-KSVD [7]	73.6	LP-β [6]	77.7
LLC [10]	73.4	FK [4]	77.8
HMP [1]	76.8	M-HMP	**82.5+0.5**

Table 1. Test accuracy on Caltech-101.

Training Images	15	30	45	60
Local NBNN [8]	33.5	40.1	/	/
LLC [10]	34.4	41.2	45.3	47.7
CRBM [9]	35.1	42.1	45.7	47.9
LP-β [6]	/	45.8	/	/
M-HMP	**42.7**	**50.7**	**54.8**	**58.0**

Table 2. Test accuracy on Caltech-256.

layer HMP networks (A1 and B1) work surprisingly well and already outperform many existing computer vision approaches, showing the benefits of learning from pixels.

They compare M-HMP with recently published state-of-the-art recognition algorithms in Tab. 1 [2]. LLC [10], LC-KSVD [7], and Asklocals [3] are one-layer sparse coding approaches. SIFT+T [5] is soft threshold coding and CRBM [9] is a convolutional variant of Restricted Boltzmann Machines (RBM). FK [4] is a Fisher Kernel based coding approach. All of them are based on SIFT. HSC [11] is a two layer sparse coding network using L1-norm regularization. Local NBNN [8] is an extension of Naive Bayesian Nearest Neighbor (NBNN). LP-β [6] is a boosting approach to combine multiple types of designed features. M-HMP achieves test accuracy superior to all of them. The average accuracy over 5 random trials in Tab. 2 [2]. They keep the same architecture as that for Caltech-101 (Section 2.1), with the only exception that the number of codewords in the final layer of HMP is increased to 2000 to accommodate for more categories and more images.

References

[1] L. Bo, X. Ren, and D. Fox. Hierarchical matching pursuit for image classification: Architecture and fast algorithms. In *NIPS*, 2011. 2

[2] L. Bo, X. Ren, and D. Fox. Multipath sparse coding using hierarchical matching pursuit. In *CVPR*, 2013. 1, 2

[3] Y. Boureau, N. Roux, F. Bach, J. Ponce, and Y. LeCun. Ask the locals: Multi-way local pooling for image recognition. In *ICCV*, 2011. 2

[4] K. Chatfield, V. Lempitsky, A. Vedaldi, and A. Zisserman. The devil is in the details: An evaluation of recent feature encoding methods. In *BMVC*, 2011. 2

[5] A. Coates and A. Ng. The importance of encoding versus training with sparse coding and vector quantization. In *ICML*, 2011. 2

[6] P. Gehler and S. Nowozin. On feature combination for multiclass object classification. In *ICCV*, 2009. 2

[7] Z. Jiang, Z. Lin, and L. Davis. Learning a discriminative dictionary for sparse coding via label consistent K-SVD. In *CVPR*, 2011. 2

[8] S. McCann and D. Lowe. Local Naive Bayes Nearest Neighbor for image classification. In *CVPR*, 2012. 2

[9] K. Sohn, D. Jung, H. Lee, and A. H. III. Efficient learning of sparse, distributed, convolutional feature representations for object recognition. In *ICCV*, 2011. 2

[10] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Guo. Locality-constrained Linear Coding for image classification. In *CVPR*, 2010. 2

[11] K. Yu, Y. Lin, and J. Lafferty. Learning image representations from the pixel level via hierarchical sparse coding. In *CVPR*, 2011. 2