Hierarchical Matching Pursuit for Image Classification: Architecture and Fast Algorithms

Liefeng Bo[1], Xiaofeng Ren[2] and Dieter Fox[1]

[1]University of Washington, [2]Intel Labs

This work

- Hierarchical matching pursuit builds a feature hierarchy layer-by-layer using an efficient matching pursuit encoder.
- **Hierarchical Matching Pursuit (HMP)**
  - ✓ Matching pursuit encoder consists of three modules: batch tree orthogonal matching pursuit, spatial pyramid max pooling, and contrast normalization;
  - ✓ Recursively run matching pursuit encoder;
  - ✓ Extract features from a typical 300×300 image in less than 1 second;
  - ✓ Outperform convolutional deep networks and SIFT based single layer sparse coding in terms of accuracy.

K-SVD (Dictionary Learning)

- K-SVD[1] learns a dictionary D and an associated sparse code matrix X from observations Y by minimizing the following reconstruction error

$$\min_{D,X} \|Y - DX\|_F^2 \quad s.t. \ \forall i, \|x_i\|_0 \le K$$

- The problem can be solved in an alternating manner. In the first stage, D is fixed and only the sparse codes are computed by orthogonal matching pursuit.

$$\min_{x_i} \|y_i - Dx_i\|^2 \quad s.t. \|x_i\|_0 \le K$$

- In the second stage, each filter in D and its associated sparse codes x are updated simultaneously by Singular Value Decomposition ($\|d_k\|_2 = 1$)

$$\|Y - DX\|_F^2 = \|Y - \sum_{j \ne k} d_j x_j^T - d_k x_k^T\|_F^2 = \|E_k - d_k x_k^T\|_F^2$$

- When the sparsity level K is set to be 1 and sparse codes are forced to be a binary(0/1), K-Means is exactly reproduced (no constraints on d_k).

[1] Aharon, Elad, and Bruckstein, IEEE Transactions on Signal Processing

Batch Tree Orthogonal Matching Pursuit

Algorithm: Batch Tree Orthogonal Matching Pursuit (BTOMP)

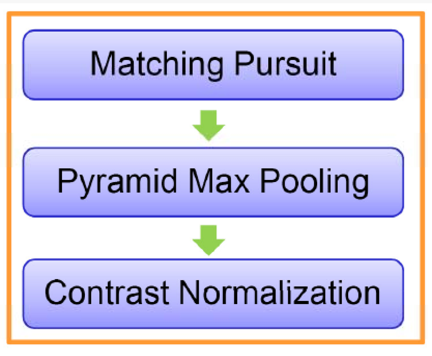
1. Input: Dictionary D, Centers C, observation y, and the desired sparsity level K
2. Output: Sparse code x such that $y \approx Dx$
3. Initialization: $I = \emptyset$, $r = y$, $\alpha = \alpha^0 = C^\top y$, $B = C^\top D$, and $x = 0$
4. For $k = 1 : K$
5. Choosing the sub-dictionary g_j: $j = \mathrm{argmax}_k |\alpha_k|$
6. Selecting the new filter: $\bar{k} = \mathrm{argmax}_{k \in g_j} |d_k^\top r|$
7. $I = I \cup \bar{k}$
8. Updating the sparse code: $x_I = (D_I^\top D_I)^{-1} D_I^\top y$
9. Updating α: $\alpha = \alpha^0 - B_I x_I$
10. Computing the residual: $r = y - D_I x_I$
11. End

- K-Means is used to group the whole dictionary into the sub-dictionaries and associate the sub-dictionaries with the learned center matrix C;
- Line 5 selects the center filter j that best matches the current residual;
- Line 6 selects the filter within the sub-dictionary associated with the center j;
- Line 8 updates sparse codes with the incremental Cholesky decomposition;
- Line 9 computes the correlation between each center and the current residual;
- If the centers C are set to be the whole dictionary, BTOMP exactly recovers the batch (exact) orthogonal matching pursuit[1].

[1] Rubinstein, Zibulevsky, and Elad,Technical report, 2008

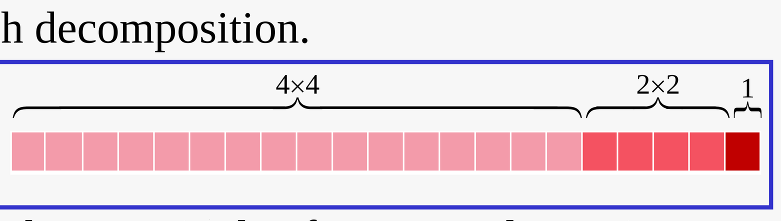
Matching Pursuit Encoder

- Matching pursuit encoder consists of three modules: BTOMP, Spatial Pyramid Max Pooling and Contrast Normalization.



- Spatial Pyramid Max Pooling aggregates the sparse codes which are spatially close, using max pooling in a multi-level patch decomposition.

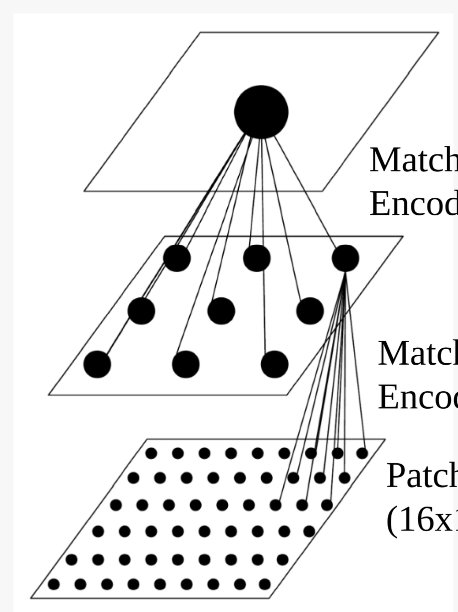
$$F(P) = \left[\max_{j \in P} |x_{j1}|, \cdots, \max_{j \in P} |x_{jm}| \right]$$



- Contrast Normalization is helpful since the magnitude of sparse codes varies significantly due to illumination and foreground-background contrast.

$$F(P) = \frac{F(P)}{\sqrt{\|F(P)\|^2 + \varepsilon}}$$

Hierarchical Matching Pursuit Encoder

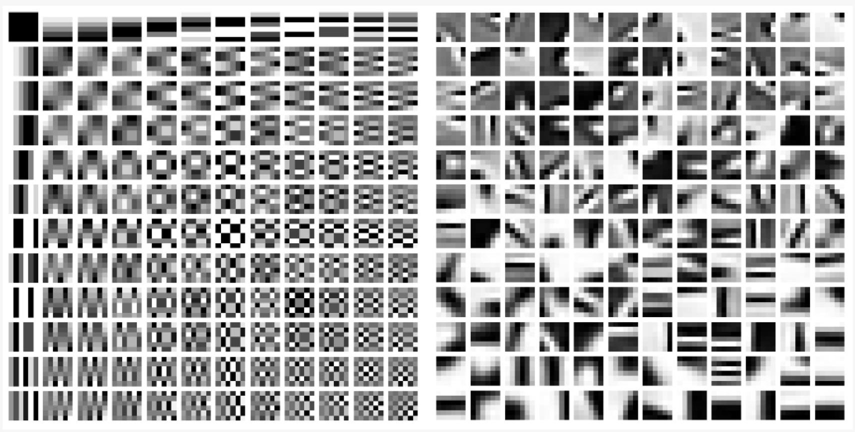


Second layer: sparse codes over patch-level features are aggregated across the whole image to produce image-level features.

First layer: sparse codes over pixel level are aggregated into patch-level features.

Object Recognition (Caltech-101)

- Dictionary is learned by K-SVD on 1,000,000 sampled patches in each layer;
- Sparsity level in the first and second layers is set to be 5 and 10, respectively;
- Dictionary size is 3 times the filter size in the first layer and 1000 in the second layer;
- Matching pursuit encoder is run on 16×16 image patches over dense grids with a step size of 4 pixels in the first layer and the whole image in the second layer;
- Train linear SVM on 30 images and test on no more than 50 images per category.
- Filters used in Discrete Cosine Transformation (DCT) and filters learned by K-SVD



- **K-SVD and DCT with different filter sizes for the first layer**

Filter size	3×3	4×4	5×5	6×6	7×7	8×8
DCT (orthogonal)	69.9	70.8	71.5	72.1	73.2	73.1
DCT (overcomplete)	69.6	71.8	73.0	74.1	73.7	73.4
K-SVD	71.8	74.4	75.9	76.8	76.3	76.1

- **Spatial Pyramid Pooling** and **Contrast Normalization** improve recognition accuracy by about 2% and 3%, respectively. Large Dictionary with 10,000 filters in the second layer is slightly better than standard setting with 1000 filters.
- **Hierarchical Matching Pursuit with K=1 (zero norm)**: about **74.0%**;
- **Running Time over a typical 300×300 image**

Algorithms	HMP (DCT)	HMP (K-SVD)	SIFT+SC	DN
Running time (Seconds)	0.4	0.8	22.4	67.5

- **Comparisons with State-of-the-art (Single Feature based Algorithms)**

	Multiple Layers					SIFT based Single Layer		
HMP	ISPD[1]	CDBN[2]	DN[3]	HSC[4]	KDES-G[5]	SPM[6]	SIFT+SP[7]	Macrofeatures[8]
76.8	65.5	65.5	66.9	74.0	75.2	64.4	73.2	75.7

[1] Kavukcuoglu, Ranzato, Fergus, and LeCun, CVPR 2009
[2] Lee, Grosse, Ranganath, and Ng, ICML 2009
[3] Zeiler, Krishnan, Taylor, and Fergus, CVPR 2010
[4] Yu, Lin, and Lafferty, CVPR 2011 (Parallel work)
[5]Bo, Ren, and Fox, NIPS 2010
[6] Lazebnik, Schmid, and Ponce, CVPR 2006
[7] Yang, Yu, Gong, and Huang, CVPR 2009
[8] Boureau, Bach, LeCun, and Ponce, CVPR 2010

Scene Recognition (MIT-Scene)

- This dataset contain 15620 images from 67 indoor scene categories;
- Train linear SVM on 80 images and test on 20 images per category;
- The experimental setting is same as with the Caltech-101 dataset except that the filter size is 4×4 (cross validation).

Algorithms	HMP	OB[1]	GIST[2]	ROI+GIST[2]	SIFT+SC
Accuracy	41.8	37.6	22.0	26.0	36.9

[1] Li, Su, Xing, and Fei-Fei., NIPS 2010
[2] Quattoni and Torralba., CVPR 2009

Event Recognition (UIUC-Sports)

- This dataset consists of 8 sport event categories with 137 to 250 images in each.
- Train linear SVM on 70 images and test on 60 images per category.
- The experimental setting is same as with the MIT-Scene dataset.

Methods	HMP	OB	SIFT+GMM [1]	SIFT+SC
Accuracy	85.7	76.3	73.4	82.7

[1] Li and Fei-Fei., ICCV 2007