GENERALIZED POOLING PYRAMID WITH HIERARCHICAL DICTIONARY SPARSE CODING FOR EVENT AND OBJECT RECOGNITION

Shuai Chen, Bo Ma *, *Pei Luo*

Beijing Laboratory of Intelligent Information Technology,Beijing Institute of Technology,China

ABSTRACT

Feature coding and vector pooling are essential for image recognition in bag-of-visual-words (BoW) method. Encoding the low-level feature to rich one and pooling it without any information loss are very challenging works. In this paper, generalized pooling pyramid with hierarchical dictionary sparse coding is introduced to get rich sparse codes and alleviate the information loss in the phase of pooling. It includes two modules: First, with the low-level feature, hierarchical dictionary is learned for sparse coding to generate the hierarchical sparse representation. Second, in the phase of vector pooling, we present generalized pooling pyramid by utilizing the probabilistic function to model the statistical distribution of sparse codes. In the generalized pooling pyramid, the Fisher vectors which are computed with Gaussian Mixture (GMM) in different levels, are fused to represent the images. The performance of our method outperforms state-of-the-art performance in a large number of image categorization experiments on the event dataset (UIUC-Sport dataset) and the object recognition dataset (Caltech101 dataset).

Index Terms— bag-of-visual-words, image recognition, sparse codes, hierarchical dictionary, generalized pooling

1. INTRODUCTION

Although, CNN model [1] show good performance than traditional BoW model [2] in image recognition in recent years. The research on selecting rich representations from low-level features is still a very challenging work with BoW model and a large number of techniques have been developed for it [3–7].

Let $\mathbf{X} = [\mathbf{x}_1, \cdots \mathbf{x}_N]^T \in \mathbb{R}^{D \times N}$ be the set of N local image descriptors with D dimensions. Yang et al. proposed linear spatial pyramid matching using sparse coding (ScSPM) [8] turn the the vector quantization (VQ) [9] into sparse coding as:

$$\begin{aligned} &\min_{\mathbf{U},\mathbf{V}} \sum_{n=1}^{N} ||\mathbf{x}_n - \mathbf{u}_n \mathbf{V}||^2 + \lambda ||\mathbf{u}_n||_1, \\ &s.t. ||\mathbf{v}_k||^2 \leq 1, k = [1, \cdots, K], \end{aligned} \tag{1}$$

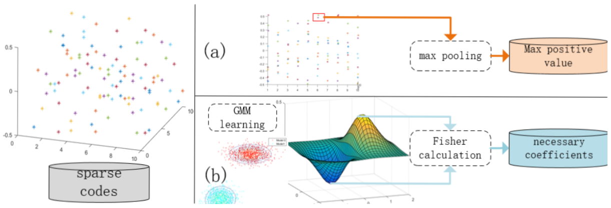


Fig. 1. Generalized pooling pyramid with hierarchical dictionary sparse coding.

where $\mathbf{V} = [\mathbf{v}_1, \cdots \mathbf{v}_K](\mathbf{v}_i \in \mathbb{R}^{D \times 1})$ is the codebook with size K, $\mathbf{U} = [\mathbf{u}_1, \cdots \mathbf{u}_N]^T(\mathbf{u}_i \in \mathbb{R}^{1 \times K})$ is the sparse code. The method encodes features by one codebook only, and then inputs the codes into the spatial pyramid, pools the sparse codes in various locations and scales with max pooling strategy to obtain the final nonlinear codes.

However, there are two major drawbacks with the sparse codes in the max pooling strategy.

A. Information loss: Max pooling is a robust pooling strategy because of choosing the maximum values which are representative in sparse codes. But, in order to achieve good performance of image recognition, negative values (or small positive values) are sometimes needed as well. In max pooling, the necessary negative coefficients (or small positive coefficients) are lost, as shown in Fig.1(a).

B. Little contribution of low levels in SPM: A large number of image categorization experiments have been done with ScSPM in our work. We find that the low levels of the pyramid contain rich spatial information, but they make little contribution to the recognition performance. This is intuitive because the codebook of each level in the spatial pyramid is the same. The pooling features generated from different levels by max pooling strategy are independent and combining the features stiffly is not well applied to image recognition with the spatial information.

For the problems,we introduced generalized pooling pyramid with hierarchical dictionary sparse coding to get rich sparse codes and alleviate the information loss in the phase of pooling. It contains two contributions:

(1) For the problem A, in order to avoid the information loss, generalized pooling pyramid strategy is introduced. The main task of the method is to establish mathematical models

*Corresponding author: bma000@bit.edu.cn (Bo Ma). This work was supported in part by the National Natural Science Foundation of China (No. 61472036).

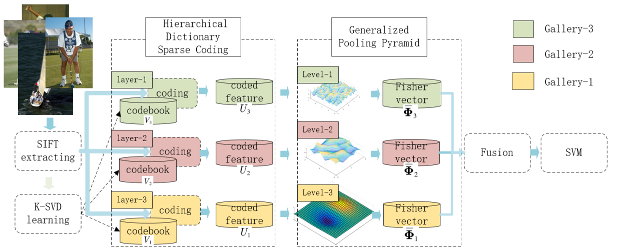


Fig. 2. Generalized pooling pyramid with hierarchical dictionary sparse coding.

by probabilistic function and model the statistical distribution of sparse codes. In our pooling strategy, the distribution of GMM is learned to calculate the Fisher vectors and the coefficients of sparse codes are weighted by it. In this way, the necessary negative coefficients (or small positive coefficients) is depended from the sparse codes, as shown in Fig.1(b).

(2) For the problem B, hierarchical dictionary sparse coding is proposed. We encode the features layer by layer with different codebook. Experiments have shown that the performance is improved when the generalized pooling pyramid is aided by the hierarchical sparse coding framework and the low levels have shown great contribution as the pooling vectors generated from different levels are fused.

2. GENERALIZED POOLING

Giving a set of D-dimensional sparse codes $\mathbf{U} = \{\mathbf{u}_i, i = 1, \cdots, N\}$ for an image sparse codes. We assume that they are generated by a parametric distribution $\Psi_\lambda(\mathbf{u})$ and independent to the corresponding codebook item. In this paper, GMM :

$$p_m(u) = \frac{1}{(2\pi)^{d/2}|\Sigma_m|^{1/2}} e^{(-\frac{1}{2}(\mathbf{u}-\boldsymbol{\mu}_m)^T \Sigma_m^{-1}(\mathbf{u}-\boldsymbol{\mu}_m))}, \quad (2)$$

is chosen as the distribution to model the generative process. p_m is the m-th component of GMM (M components), we design different M for different dataset.

Then, $\Psi_\lambda(\mathbf{u})$ can be written as:

$$\Psi_\lambda(\mathbf{u}) = \sum_{m=1}^{M} \omega_m p_m(\mathbf{u}), \quad (3)$$

where the parameter $\lambda = \{\omega_m, \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m\}_{m=1,\cdots,M}$ is calculated by GMM training with sparse codes. To ensure the distribution for $\Psi_\lambda(\mathbf{u})$ to be valid, each ω_m must be equal to or greater than zero and the sum of all ω_m is to be one. we re-parameterizing ω_m as

$$\omega_m = \frac{e^{\alpha_m}}{\Sigma_{i=1}^{M} e^{\alpha_i}}, \quad (4)$$

where α_i is used to avoid enforcing explicitly the constraints in Eqn. (3).

Furthermore, we assume $\boldsymbol{\delta}_m$ is diagonal covariance matrix [6] for each Gaussian component. The Fisher vector will calculate the sum of gradient statistic with the parameter of GMM by the following:

$$\begin{aligned}
\Theta_{\alpha_m} &= \tfrac{1}{\sqrt{\omega_m}} \sum_{i=1}^{N} (\Upsilon_i(m) - \omega_m), \\
\Theta_{\boldsymbol{\mu}_m} &= \tfrac{1}{\sqrt{\omega_m}} \sum_{i=1}^{N} \Upsilon_i(m)(\tfrac{\mathbf{u}_i - \boldsymbol{\mu}_m}{\boldsymbol{\delta}_m}), \\
\Theta_{\boldsymbol{\delta}_m} &= \tfrac{1}{\sqrt{\omega_m}} \sum_{i=1}^{N} \Upsilon_i(m) \tfrac{1}{\sqrt{2}} (\tfrac{(\mathbf{u}_i - \boldsymbol{\mu}_m)^2}{{\boldsymbol{\delta}_m}^2} - 1).
\end{aligned} \quad (5)$$

The gradient is generated by derivation [10]. Θ_{α_m}, $\Theta_{\boldsymbol{\mu}_m}$ and $\Theta_{\boldsymbol{\delta}_m}$ are normalized gradients of sparse features with respect to α_m, $\boldsymbol{\mu}_m$ and $\boldsymbol{\delta}_m$.

$$\Upsilon_i(m) = \frac{\omega_m p_m(\mathbf{u}_i)}{\sum_{j=1}^{M} \omega_j p_j(\mathbf{u}_i)} \quad (6)$$

is the soft assignment of $\mathbf{u}_i$ for the m-th component of GMM.

Considering the gradient with respect to the weight parameter Θ_{α_m} brings little additional information, the Fisher vector is obtained by concatenating the gradients , $\Theta_{\boldsymbol{\mu}_m}$,$\Theta_{\boldsymbol{\delta}_m}$ only. Therefore, the dimension of the Fisher vector $\mathbf{\Phi} = \{\Theta_{\boldsymbol{\mu}_m}, \Theta_{\boldsymbol{\delta}_m}\}$ is $2dM$ in our method. In addition, the L2 normalization is necessary for Fisher vectors to obtain competitive results:

$$\bar{\mathbf{\Phi}} = \frac{\mathbf{\Phi}}{\sqrt{||\mathbf{\Phi}||_2^2 + \varepsilon}}, \quad (7)$$

ε is a small positive number, where a small threshold is able to make low contrast patches more separate from high contrast image patches [11].

3. HIERARCHICAL DICTIONARY SPARSE CODING

Hierarchical dictionary sparse coding is divided into two parts: the stage of training and the stage of coding.

In the stage of training, K-SVD [12]:

$$\begin{aligned}
&\{\mathbf{V}_l\} \leftarrow \min_{\mathbf{u}} \sum_{n=1}^{N} ||\mathbf{x}_n - \sum_{k=1}^{PY*K} \mathbf{u}_{n,k} \mathbf{v}^{(l)}{}_k||^2, \\
&s.t. ||\mathbf{v}^{(l)}{}_k||^2 \leq 1, \forall k.
\end{aligned} \quad (8)$$

is used to generate the dictionaries, where K denotes the codebook size in level 1 of the generalized pooling pyramid,$\mathbf{V}_l = [\mathbf{v}_1^{(l)}, \cdots \mathbf{v}_{PY*K}^{(l)}](\mathbf{v}_i^{(l)} \in \mathbb{R}^{D \times 1})$ is the codebook, $\mathbf{X} = [\mathbf{x}_1, \cdots \mathbf{x}_N]^T \in \mathbb{R}^{D \times N}$ be the set of N local image descriptors with D dimensions.

$$PY = \frac{1}{2^{l-1}}, \quad (9)$$

where $l = [1, 2, \cdots L]$ is corresponding to level l of the generalized pooling pyramid respectively and L is the level of

Algorithm 1 Hierarchical Dictionary Sparse Coding with Generalized Pooling Pyramid

```
Input: SIFT descriptors: $\mathbf{X}$;
Ouput: The representations $\{\overline{\mathbf{\Phi}}^{(t)}\}_{t=1}^{T}$ generated by Generalized Pooling Pyramid with Hierarchical Dictionary Sparse Coding
$L \leftarrow 3, PY \leftarrow \frac{1}{2}, \varepsilon \leftarrow 0.01$
$Calculate\ sparse\ codes\ by\ Hierarchical\ Dictionary$
for $l = 1 \rightarrow L$ do
    $Optimize\ \{\mathbf{V}_l\}\ of\ K-SVD\ Eqn.\ (9)$
    $Fixed\ \{\mathbf{V}_l\}$
    $Optimize\ \{\mathbf{U}_l\}\ of\ sparse\ coding\ Eqn.(11)$
end for
$Generate\ Fisher\ vector\ \mathbf{\Phi}\ in\ every\ level$
for $l = 1 \rightarrow L$ do
    $Choose\ suit\ GMM\ numbers$
    $Calculate\ parameter\ of\ GMM\ in\ every\ level$
    $\{\omega_m, \boldsymbol{\mu}_m, \Sigma_m\}_l \leftarrow GMM\ training$
    $Generate\ Fisher\ vector\ F\ in\ every\ level$
    $\mathbf{\Phi}\{(\vartheta_{\alpha_k}), (\vartheta_{\boldsymbol{\mu}_k})\} \leftarrow\ calculate\ Eqn.(5)$
    $\overline{\mathbf{\Phi}} \leftarrow \frac{\mathbf{\Phi}}{\sqrt{||\mathbf{\Phi}||_2^2+\varepsilon}}$
end for
```

spatial pyramid model. Sequentially, the hierarchical codebook

$$\mathbf{V}_{HY} = \{\mathbf{V}_1, \cdots, \mathbf{V}_l\}, \mathbf{V}_l \in \mathbb{R}^{PY*K \times D}, \tag{10}$$

is calculated for every level of the spatial pyramid model.

In the the stage of coding, hierarchical codebook $\mathbf{V}_{HY}$ is fixed to encode the features:

$$\{\mathbf{U}_l\} \leftarrow \min_{\mathbf{v}} \sum_{n=1}^{N} ||\mathbf{x}_n - \sum_{k=1}^{PY*K} \mathbf{u}_{n,k}^{(l)} \mathbf{v}_k||^2 + \sum_{n=1}^{N} \sum_{k=1}^{PY*K} |\mathbf{u}_{n,k}^{(l)}|, \tag{11}$$

where $\mathbf{U}_l = [\mathbf{u}_1^{(l)}, \cdots \mathbf{u}_N^{(l)}]^T (\mathbf{u}_i^{(l)} \in \mathbb{R}^{1 \times PY*K})$ denote the sparse codes. Then the hierarchical sparse code:

$$\mathbf{U}_{HY} = \{\mathbf{U}_1, \cdots, \mathbf{U}_l\}, \mathbf{U}_l \in \mathbb{R}^{N \times PY*K}, \tag{12}$$

is generated.

4. OVERVIEW ON OUR RECOGNITION METHOD

As show in Fig.2, the hierarchical dictionary $\mathbf{V}_{PY}$ is calculated by the SIFT [13] features, and then the hierarchical sparse codes $\mathbf{U}_{PY}$ are encoded with the hierarchical dictionary. In generalized pooling pyramid, every level is corresponding to one layer of hierarchical sparse coding, namely, $\mathbf{V}_l \in \mathbb{R}^{D \times \frac{K}{2^{(l-1)}}}$ and $\mathbf{U}_l \in \mathbb{R}^{N \times \frac{K}{2^{(l-1)}}}$ are corresponding to level l of the generalized pooling pyramid. The level of spatial pyramid model is fixed as L= 3 and the number of total grid cell is $\sum_{l=1}^{3} 2^{2(l-1)} = 21$ [14]. Finally, a linear SVM is fed with the fused Fisher vectors generated from different levels as our classifier.

For the convenience of description, the structure of layer l codebook and sparse coding with level l generalized pooling pyramid is named as gallery l. As shown in Fig. 2.

Table 1. The effectiveness of hierarchical structure on the Caltech-101 dataset.

method	size	Gallery1	Gallery2	Gallery3	Fusion
		(2GMM)	(1GMM)	(1GMM)	(best)
Ours	2048	0.612	0.765	**0.819**	**0.841**
	4096	0.643	**0.787**	0.803	±
	8192	**0.712**	0.724	0.782	**0.131**
		level-1	level-2	level-3	
ScSPM	1024	0.458	0.631	0.729	0.732
HScSPM	2048	0.545	0.695	**0.751**	
	4096	0.624	**0.727**	0.723	**0.775**
	8192	**0.691**	0.704	0.690	

Table 2. The effectiveness of hierarchical structure on UIUC-Sport dataset.

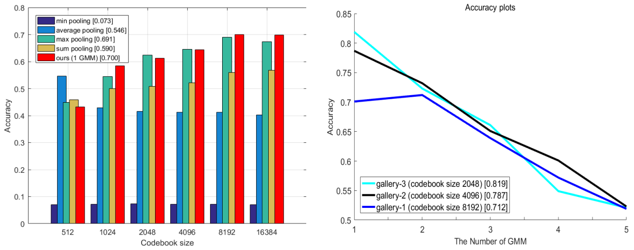
method	size	Gallery1	Gallery2	Gallery3	Fusion
		(2GMM)	(1GMM)	(1GMM)	
Ours	2048	0.830	0.845	**0.889**	**0.899**
	4096	0.848	**0.877**	0.803	±
	8192	**0.868**	0.844	0.792	**0.165**
		level-1	level-2	level-3	
ScSPM	1024	0.784	0.760	0.804	0.827
HScSPM	2048	0.817	0.785	**0.829**	
	4096	0.839	**0.767**	0.813	0.864
	8192	**0.840**	0.744	0.792	

5. EXPERIMENTS

We compare our method with many state-of-the-art algorithms on two recognition tasks: Object recognition (Caltech101 [15]) and Event recognition (UIUC-Sports [16]). To extract the SIFT features, we make the image to 16 × 16 pixels patches with 8 pixels overlapping and the size of the local SIFT descriptor is 128. All images are transformed into grayscale with normalized to [0, 1] and resized to be no larger than 300 × 300 pixels. ε in the normalization is 0.01.

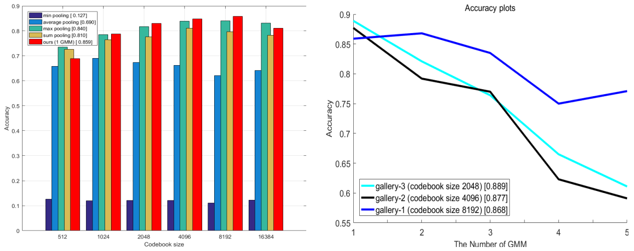
First, we demonstrate that the generalized pooling strategy is doing well than other pooling strategies (i.e., average, max and min, summation pooling) in recognition tasks. To keep things fair, we use gallery 1 (1 GMM without hierarchical structure) only to show the comparisons under the same protocols and the influence of the codebook size will be shown.

Second, We investigate how the hierarchical structure can help improve the task. we will make comparisons of



(a) (b)

Fig. 3. (caltech-101) Different pooling representations under gallery 1 with different size of codebook and 1GMM in(a). (b):The effectiveness of the GMM number on different gallery with the best size codebook.



(a) (b)

Fig. 4. (UIUC-SPORT) Different pooling representations under gallery 1 with different size of codebook and 1GMM in(a).The effectiveness of the GMM number on different gallery with the best size codebook.

Table 3. Performance comparison $(\%)$on the caltech-101

Algorithms	Accuracy
[8]ScSPM	73.20 ± 0.54
[18]LLC	73.44
[19]LR-Sc^{+}SPM	75.68 ± 0.89
[11]MHP	76.8
[20]FK	77.8
[21]H-Tom	77.78 ± 0.86
[22]M-MHP	82.5 ± 0.5
HScSPM	77.5 ± 0.5
Ours	**84.10** ± **1.31**

Table 4. Performance comparison$(\%)$on the UIUC-SPORT

Algorithms	Accuracy
[8]ScSPM	82.74 ± 1.46
[23]LScSPM	85.31 ± 0.51
[19]Sc^{+}SPM	83.77 ± 0.97
[19]LR-Sc^{+}SPM	86.69 ± 1.66
[24]SVC	87.26 ± 0.81
[25]IFK	88.80 ± 0.81
HScSPM	86.40 ± 0.81
Ours	**89.87** ± **1.65**

our method (GMMs with hierarchical structure) with ScSPM (without hierarchical structure) and HScSPM (ScSPM with hierarchical structure). Furthermore, the best cookbook size and GMM numbers for our method will be found. In the end, we will evaluate our method with many state-of-the-art recognition algorithms under the same protocol.

In our experiment, the process is repeated for 10 times with selecting the training set randomly and the average of the correct classification rates is used as the recognition rate. In the classification stage, LIBSVM [17] which adopts the SVM with linear kernel, is used as the classifier. All of the experiments are implemented in Matlab with CPU Intel Core i7 and the memory is 32G.

5.1. Object recognition: Caltech-101

We follow the common experiment setup as did in [8, 18, 19, 26, 27] and randomly choose 30 images per category for training and the rest for testing. Figure 3(a) shows the performance comparison of our pooling method with other pooling strategy. TABLE 1 shows the effectiveness of hierarchical structure. And as show in Figure 3(b), the best setting is: gallery 1 with 8196 codebook size and 2 GMM, gallery 2 with 4096 codebook size and 1 GMM, gallery 3 with 2048 codebook size and 1 GMM. Finally, the fused feature with the best setting gets the state-of-the-art performance **84.10** ± **1.31**,as shown in TABLE 3.

5.2. Event recognition: UIUC-Sport

We randomly select 70 images from each class for training [8, 28] and test on the rest images. We compare our method under gallery 1 with other pooling methods in Figure 4(a). Moreover, the effectiveness of hierarchical structure is shown in TABLE 4. Combining Figure 4 (b) the results show the best setting, Gallery 1: 2 GMM, codebook size 8192; Gallery 2: 1 GMM, codebook size 4096; Gallery 3: 1 GMM, codebook size 2048. Under the setting,TABLE 4 shows our method get **89.87** ± **1.65** comparison of our method with several other methods [8, 19, 23–25, 28] on the UIUC-Sport dataset.

6. CONCLUSIONS

We introduce a framework which learns meaningful multi-gallery united representations from images gallery by gallery. Hierarchical dictionary sparse coding is introduced to encode the feature layer by layer, and then input them to the corresponding level in generalized pooling pyramid. Further,we use GMM as the probabilistic function to model the statistical distribution of sparse codes, and then Fisher vector is used to improve the recognition performance. And the combined feature shows the high recognition rate on several benchmarks.

7. REFERENCES

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *NIPS (2012)*, pp. 1097–1105.

[2] G. Csurka, C. Dance, L. Fan, J. Willamowski, and Bray, "Visual categorization with bags of keypoints," in *ECCV (2004)*, pp. 1–2.

[3] R. Gopalan, "Hierarchical sparse coding with geometric prior for visual geo-location," in *CVPR (2015)*, pp. 2432–2439.

[4] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[5] G.-S. Xie, X.-Y. Zhang, X. Shu, S. Yan, and C.-L. Liu, "Task-driven feature pooling for image classification," in *ICCV (2015)*, pp. 1179–1187.

[6] B. Ma, H. Hu, J. Shen, Y. Liu, and L. Shao, "Generalized pooling for robust object tracking," *TIP (2016)*, pp. 4199–4208.

[7] H. Luo and H. Lu, "Multi-level sparse coding for human action recognition," in *IHMSC (2016)*, vol. 1, pp. 460–463.

[8] J. Yang, K. Yu, Y. Gong, and T. Huang, "Linear spatial pyramid matching using sparse coding for image classification," in *CVPR (2009)*, pp. 1794–1801.

[9] A. Gersho and R. M. Gray, *Vector quantization and signal compression*. Springer Science & Business Media, 2012, vol. 159.

[10] J. Sanchez, F. Perronnin, T. Mensink, and J. Verbeek, "Image classification with the fisher vector: Theory and practice," *IJCV (2013)*, pp. 222–245.

[11] L. Bo, X. Ren, and D. Fox, "Hierarchical matching pursuit for image classification: Architecture and fast algorithms," in *NIPS (2011)*, pp. 2115–2123.

[12] M. Aharon, M. Elad, and A. Bruckstein, "Svd: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Transactions on signal processing (2006)*, pp. 4311–4322.

[13] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *IJCV (2004)*, pp. 91–110.

[14] S. Lazebnik, C. Schmid, J. Ponce *et al.*, "Spatial pyramid matching," *Object Categorization: Computer and Human Vision Perspectives (2009)*.

[15] L. Fei-Fei, R. Fergus, and P. Perona, "Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories," *CVIU (2007)*, pp. 59–70.

[16] Y. Yang and S. Newsam, "Bag-of-visual-words and spatial extensions for land-use classification," in *Proceedings of the 18th SIGSPATIAL international conference on advances in geographic information systems (2010)*, pp. 270–279.

[17] C.-C. Chang and C.-J. Lin, "Libsvm: a library for support vector machines," *TIST (2011)*, vol. 2, no. 3, p. 27.

[18] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong, "Locality-constrained linear coding for image classification," in *CVPR (2010)*. IEEE, 2010, pp. 3360–3367.

[19] C. Zhang, J. Liu, Q. Tian, C. Xu, H. Lu, and S. Ma, "Image classification by non-negative sparse coding, low-rank and sparse decomposition," in *CVPR (2011)*, pp. 1673–1680.

[20] K. Chatfield, V. S. Lempitsky, A. Vedaldi, and A. Zisserman, "The devil is in the details: an evaluation of recent feature encoding methods." in *BMVC (2011)*, p. 8.

[21] K. Huang, C. Wang, and D. Tao, "High-order topology modeling of visual words for image classification," *TIP (2015)*, pp. 3598–3608.

[22] L. Bo, X. Ren, and D. Fox, "Multipath sparse coding using hierarchical matching pursuit," in *CVPR (2013)*, pp. 660–667.

[23] S. Gao, I. W.-H. Tsang, L.-T. Chia, and P. Zhao, "Local features are not lonely–laplacian sparse coding for image classification," in *CVPR (2010)*, pp. 3555–3561.

[24] F. Perronnin, J. Snchez, and T. Mensink, "Improving the fisher kernel for large-scale image classification," in *ECCV (2010)*, pp. 143–156.

[25] X. Zhou, K. Yu, T. Zhang, and T. S. Huang, "Image classification using super-vector coding of local image descriptors," in *ECCV (2010)*, pp. 141–154.

[26] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *CVPR (2006)*, pp. 2169–2178.

[27] J. C. Van Gemert, C. J. Veenman, A. W. Smeulders, and Geusebroek, "Visual word ambiguity," *TPAMI (2010)*, pp. 1271–1283.

[28] J. Wu and J. M. Rehg, "Beyond the euclidean distance: Creating effective visual codebooks using the histogram intersection kernel," in *ICCV (2009)*, pp. 630–637.