Sparse Coding for Histograms of Local Binary Patterns Applied for Image Categorization: Toward a Bag-of-Scenes Analysis

Sébastien PARIS[1], Xanadu HALKIAS[2] and Hervé GLOTIN[2]
[1]*DYNI team, LSIS CNRS UMR 7296, Aix-Marseille University*
[2]*DYNI team, LSIS CNRS UMR 7296, Université Sud Toulon-Var*
sebastien.paris@lsis.org, halkias@univ-tln.fr, glotin@univ-tln.fr

Abstract

In this work[1], we propose a novel approach for image categorization, which we will refer to as Bag-of-Scenes (BoS). It is based on the association of Sparse coding (Sc) and pooling techniques applied to histograms of multi-scale Local Binary Patterns (LBP) and its improved variant. This approach can be considered as a 2-layer hierarchical architecture. The first layer, encodes general local patch's structure via histograms of LBP, and the second, encodes the relationships between pre-analyzed LBP-scenes. Our method outperforms SIFT-based approaches using Sc techniques and can be trained efficiently with a simple linear SVM. Our BoS method achieves 87.02%*,* 87.71% *and* 79.05% *of accuracy for Scene-15, UIUC-Sport and Caltech101 datasets respectively.*

1. Introduction

Image categorization consists of assigning a unique label with a generally high-level semantic value to an image. This has long been a challenging problem area in both computer vision and robotics and can mainly be viewed as belonging to the broader supervised classification framework. The difficulty of the task can be partly explained by the high-dimensional input space of the images as well as the high-level semantic visual concepts that lead to large intra-class variation.

The most common, *direct* framework in current vision systems is to extract directly from the images meaningful features (using shape/texture/color information) in order to achieve the maximum generalization capacity during the classification stage. Examples of such popular features in computer vision and human cognition inspired models include GIST [14] based on a bank of Gabor filters and the HOG descriptors [5].

Widely used in face recognition [21] and scene categorization [7, 16, 20], Histograms of LBP (HLBP) [13], are competitive features that achieve state-of-the-art performances in the tasks at hand. Each LBP can be considered as a non-parametric, local, visual micro-pattern texture encoding, mainly, contours and differential excitation information of the 8 neighbors surrounding a central pixel [8]. The total number of different LBPs is relatively small and by construction it is finite (from 256 up to 512). HLBP, which counts the occurrence of each LBP in the scene, can easily capture general structures in the visual scene by integrating information in a Region of Interest (ROI), while being less sensitive to local high frequency details. This property is important when the desire is to generalize visual concepts. As depicted in this work, it is advantageous to extend this analysis for several sizes of local ROIs using a spatial pyramid denoted by $\underline{\mathbf{\Lambda}}$.

Recently, the alternative scheme of *Bag-of-Features* (BoF) has been employed in several computer vision tasks with wide success. It offers a deeper extraction of visual concepts and improves accuracy of computer vision systems. BoF image representation [9] shares the same idea as HLBP: counting the presence (or combination) of visual patterns in the scene. BoF contains at least three modules prior to the classification stage: (i) region selection for patch extraction; (ii) codebook/dictionary generation and feature quantization; (iii) frequency histogram based image representation with spatial pyramidal matching (SPM).

In general, SIFT/HOG patches [11, 5] are employed in the first module or more recently with the efficient but computationally expensive Kernel descriptors (see [2]). These visual descriptors are then encoded, in an unsupervised manner, into a moderate sized dictionary using Vector Quantization (VQ) (see [9]). In [18], Wu and *al* were first to introduce LBP (*via* CENTRIST) into BoF

[1]Granded by COGNILEGO ANR 2010-CORD-013 and PEPS RUPTURE Scale Swarm Vision

framework coupled with histogram intersection kernel (HIK). In order to improve encoding scheme, it has been shown that local-constrained linear coding (LLC) [15], orthogonal matching pursuit (OMP) [1] or Sparse coding (Sc) [19, 7] can easily be plugged into this BoF framework as a replacement for VQ. Moreover, pooling techniques coupled with a second SPM [9] (denoted by $\overline{\mathbf{\Lambda}}$) can be effectively used as a replacement for the global histogram based image representation.

In this paper we first re-introduce two multi-scale variants of the LBP operator coupled with a spatial pyramid $\underline{\mathbf{\Lambda}}$ analysis[2](generalizing the framework of macro-features of Boureau and *al* [3]). We propose the use HLBP into the Sc framework and call *Bag-of-Scenes* (BoS), this new approach for scene categorization. The novel obtained feature can be trained efficiently with linear large-scale classifier. BoS can be seen as a two layer Hierarchical BoF analysis: a first fast parametric contractive low-dimension manifold encoder *via* HLBP and a second high-dimension sparse encoder *via* Sc.

2 Multi-Scale Histogram of LBP

We present two existing multi-scale versions of the LBP [10] operator for an image/patch $\boldsymbol{I}$ $(n_y \times n_x)$, *i.e.* the C operator and its *improved* variant IC. Basically operator C encodes the relationship between a central block of $(s \times s)$ pixels located in (y_c, x_c) with its 8 neighboring blocks, whereas IC adds a ninth bit encoding a term homogeneous to the differential excitation. Both can be considered as a parametric local texture encoder for scale s. In order to capture information at different scales, the range analysis $s \in \mathcal{S}$, is typically set at $\mathcal{S} = [1, 2, 3, 4]$ for this paper, where $S = Card(\mathcal{S})$. These 2 micro-codes are defined as follows:

$$\begin{cases} C(y_c, x_c, s) = \sum_{i=0}^{i=7} 2^i \mathbb{1}_{\{A_i \geq A_c\}} \\ IC(y_c, x_c, s) = \sum_{i=0}^{i=7} 2^i \mathbb{1}_{\{A_i \geq A_c\}} + 2^8 \mathbb{1}_{\left\{\sum_{i=0}^{7} A_i \geq 8Ac\right\}} \end{cases}. \quad (1)$$

The different areas $\{A_i\}$ and A_c in eq.(1) can be computed efficiently using the image integral technique. Efficient descriptors corresponding to the operator $op = C$ or $op = IC$, are obtained by counting occurrences of the j^{th} parametric visual LBP at scale s in a ROI $\underline{\boldsymbol{R}} \subseteq \boldsymbol{I}$:

$$h_{op}(\underline{\boldsymbol{R}}, j, s) = \sum_{(x_c, y_c) \in \underline{\boldsymbol{R}}} \mathbb{1}_{\{op(y_c, x_c, s) = j\}}.$$

[2]All ROI's notations for the first layer will be underlined, whereas ROI's notations for the second layer will be overlined.

The full histogram HC for $op = C$ (respectively HIC for $op = IC$), with $b = 256$ bins (512 respectively), is defined by: $\boldsymbol{h}_{op}(\underline{\boldsymbol{R}}, s) \triangleq [h_{op}(\underline{\boldsymbol{R}}, 0, s), \ldots, h_{op}(\underline{\boldsymbol{R}}, b-1, s)]$.

2.1 HC/HIC coupled with Spatial Pyramid $\underline{\mathbf{\Lambda}}$

Instead of using image/patch $\boldsymbol{I}$ as the only ROI, the entire zone can be divided into several sub-windows (possibly overlapping) *via* a spatial pyramid $\underline{\mathbf{\Lambda}}$ defined with $\underline{L}$ layers. For each layer $l = 0, \ldots, \underline{L} - 1$, $\boldsymbol{I}$ is divided in $\{\underline{\boldsymbol{R}}_{l,v}\}$ ROI's, with $v = 0, \ldots, \underline{V}_l - 1$ where $\underline{V}_l$ denotes the total number of sub-windows for the l^{th} layer. A total of $\underline{V} = \sum_{l=0}^{\underline{L}-1} \underline{V}_l$ histograms $\boldsymbol{h}_{op}(\underline{\boldsymbol{R}}_{l,v}, s)$ are computed where $\underline{\boldsymbol{R}}_{l,v}$ is the v^{th} sub-window of layer l. For each scale s, the vector $\boldsymbol{x}(\underline{\mathbf{\Lambda}}, s)$ is obtained by the weighted concatenation of all sub-window histograms such as: $\boldsymbol{x}(\underline{\mathbf{\Lambda}}, s) \triangleq \{\underline{\lambda}_l \boldsymbol{h}_{op}(\underline{\boldsymbol{R}}_{l,v}, s)\}$, where $l = 0, \ldots, \underline{L} - 1$, $v = 0, \ldots, \underline{V}_l - 1$ and $\underline{\lambda}_l$ denotes the weight applied to all sub-windows of the l^{th} layer.

3 Sparse Coding on HC/HIC patches

Here, we replace the collection of usual SIFT patches located in $\{\boldsymbol{O}_k \subseteq \boldsymbol{I}\}$ densely sampled on a grid over the entire image $\boldsymbol{I}$ by our HC/HIC local descriptor $\boldsymbol{x}(\underline{\mathbf{\Lambda}}, s)$ seen previously. Specifically, $\forall s \in \mathcal{S}$, F patches of size $(m \times m)$ associated with ROI's $\{\boldsymbol{O}_k\}$ (possibly overlapping) are extracted for $k = 0, \ldots, F-1$. For a complete dataset containing N images and $\forall s \in \mathcal{S}$, we obtain a collection of $P = TS$ patches $\boldsymbol{X} \triangleq \{\boldsymbol{x}_i\}$, $i = 1, \ldots, P$, where $T = NF$. We define by $\boldsymbol{X}(s) \subseteq \boldsymbol{X}$, the subset of patches $\boldsymbol{x}_i$ at scale s with T elements.

3.1 Sparse coding overview

In order to obtain highly discriminative visual features, a common procedure consists of encoding each patch $\boldsymbol{x}_i \in \boldsymbol{X}(s)$ at scale s through an unsupervised trained dictionary $\boldsymbol{D} \triangleq [\boldsymbol{d}_1, \ldots, \boldsymbol{d}_K] \in \mathbb{R}^{d \times K}$, where K denotes the number of dictionary elements, and its corresponding weight vector $\boldsymbol{c}_i \in \mathbb{R}^K$. In the Sc approach, in order to i) reduce the quantization error and ii) to have a more realistic representation of the patches, each vector $\boldsymbol{x}_i$ is now expressed as a linear combination of a few vectors of the dictionary $\boldsymbol{D}$ and not only by a single one. The problem is formulated using the following equation:

$$\arg\min_{\boldsymbol{D}, \boldsymbol{C}} \sum_{i=1}^{T} \|\boldsymbol{x}_i - \boldsymbol{D}\boldsymbol{c}_i\|_2^2 + \beta \|\boldsymbol{c}_i\|_{\ell_1} \quad s.t. \quad \|\boldsymbol{c}_i\|_{\ell_1} = 1,$$

where the sparsity in controlled by the parameter β. The last equation is not jointly convex in $(\boldsymbol{D}, \boldsymbol{C})$ and a common procedure consists of optimizing alternatively $\boldsymbol{D}$ given $\boldsymbol{C}$ by a block coordinate descent and then $\boldsymbol{C}$ given $\boldsymbol{D}$ by a LASSO procedure [12]. At the end of the process, for $\forall s \in \mathcal{S}$, a trained dictionary $\widehat{\boldsymbol{D}}(s)$ is obtained.

3.2 Max pooling and SPM $\overline{\boldsymbol{\Lambda}}$: HC-ScSPM/HIC-ScSPM

For an image $\boldsymbol{I}$ of the dataset and given $\widehat{\boldsymbol{D}}(s)$ at scale s, F sparse vectors $\{\boldsymbol{c}_k(s)\}$ are computed by a LASSO algorithm [12]. An efficient descriptor $\boldsymbol{z}(s) \triangleq \left[z^0(s), \ldots, z^{K-1}(s)\right] \in \mathbb{R}^K$ is obtained by the following max-pooling procedure [19, 3]:

$$z^j(s) \triangleq \max_{k|\boldsymbol{x}_k \in \overline{\boldsymbol{R}}} (|c_k^j(s)|),\ j = 0, \ldots, K-1, \quad (2)$$

where each element of $\boldsymbol{z}(s)$ represents the max-response of the absolute value of sparse codes belonging to the ROI $\overline{\boldsymbol{R}}$. In order to improve accuracy, as in section 2.1, a spatial pyramidal matching procedure helps to perform a more robust local analysis. The spatial pyramid $\overline{\boldsymbol{\Lambda}}$ has $\overline{V} = \sum_{l=0}^{\overline{L}-1} \overline{V}_l$ ROIs $\{\overline{\boldsymbol{R}}_{l,v}\}$ with $l = 0, \ldots, \overline{L}-1$, $v = 0, \ldots, \overline{V}_l - 1$. The quantity $\boldsymbol{z}_{l,v}^j(s)$ for each ROI $\overline{\boldsymbol{R}}_{l,v}$ is computed by: $z_{l,v}^j(s) \triangleq \max_{k|\boldsymbol{x}_k \in \overline{\boldsymbol{R}}_{l,v}} (|c_k^j(s)|)$.

The final descriptor $\boldsymbol{z}(\overline{\boldsymbol{\Lambda}}) \in \mathbb{R}^{\overline{d}}$, denoted by HC-ScSPM/HIC-ScSPM respectively, where $\overline{d} = K\overline{V}S$, will be defined by the weighted concatenation of all the $\boldsymbol{z}_{l,v}(s)$ vectors: $\boldsymbol{z}(\overline{\boldsymbol{\Lambda}}) \triangleq \{\overline{\lambda}_l \boldsymbol{z}_{l,v}(s)\}$. $\boldsymbol{z}(\overline{\boldsymbol{\Lambda}})$ is then ℓ_2 normalized.

4 Results and conclusion

We test our approach on Scene-15 [9], UIUC-Sport [7] and Caltech101 [1] datasets. We define the second layer SPM matrix $\overline{\boldsymbol{\Lambda}}$ with $\overline{L}$ levels by $\overline{\boldsymbol{\Lambda}} \triangleq \left[\overline{\boldsymbol{r}}_y, \overline{\boldsymbol{r}}_x, \overline{\boldsymbol{d}}_y, \overline{\boldsymbol{d}}_x, \overline{\boldsymbol{\lambda}}\right]$ of size $(\overline{L} \times 5)$. For a level $l \in \{0, \ldots, \overline{L}-1\}$, the image $\boldsymbol{I}$ is divided into potentially overlapping sub-windows $\overline{\boldsymbol{R}}_{l,v}$ of size $(\overline{h}_l \times \overline{w}_l)$ and its associated weight is $\overline{\lambda}_l$. In our implementation, $\overline{h}_l \triangleq \lfloor n_y.\overline{r}_{y,l} \rfloor$ and $\overline{w}_l \triangleq \lfloor n_x.\overline{r}_{x,l} \rfloor$ where $\overline{r}_{y,l}$, $\overline{r}_{x,l}$ and $\overline{\lambda}_l$ are the l^{th} element of vectors $\overline{\boldsymbol{r}}_y$, $\overline{\boldsymbol{r}}_x$ and $\overline{\boldsymbol{\lambda}}$ respectively. Sub-window shifts in x-y axis are defined by integers $\overline{\delta}_{y,l} \triangleq \lfloor n_y.\overline{d}_{y,l} \rfloor$ and $\overline{\delta}_{x,l} \triangleq \lfloor n_x.\overline{d}_{x,l} \rfloor$ where $\overline{d}_{y,l}$ and $\overline{d}_{x,l}$ are elements of $\overline{\boldsymbol{d}}_y$ and $\overline{\boldsymbol{d}}_x$ respectively. The total number of sub-windows is equal to: $\overline{V} = \sum_{l=0}^{\overline{L}-1} \overline{V}_l = \sum_{l=0}^{\overline{L}-1} \lfloor \frac{(1-\overline{r}_{y,l})}{\overline{d}_{y,l}} + 1 \rfloor . \lfloor \frac{(1-\overline{r}_{x,l})}{\overline{d}_{x,l}} + 1 \rfloor$.

For the first layer intra-patch SPM matrix $\underline{\boldsymbol{\Lambda}}$, we remplace in the previous definition $n_y = n_x = m$. For all datasets, we will particularize $\underline{\boldsymbol{\Lambda}}_1 = [\ 1\ \ 1\ \ 1\ \ 1\ \ 1\ ]$, $\underline{\boldsymbol{\Lambda}}_2 = [\ \frac{1}{2}\ \ \frac{1}{2}\ \ \frac{1}{2}\ \ \frac{1}{2}\ \ 1\ ]$, $\overline{\boldsymbol{\Lambda}}_1 = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{6} & \frac{1}{6} & 1 \end{bmatrix}$ and $\overline{\boldsymbol{\Lambda}}_2 = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ \frac{1}{2} & \frac{1}{2} & \frac{1}{2} & \frac{1}{2} & 1 \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & 1 \end{bmatrix}$ leading to $\underline{L}_1 = 1, \underline{V}_1 = 1$, $\underline{L}_2 = 2, \underline{V}_2 = 4$, $\overline{L}_1 = 2, \overline{V}_1 = 26$ and $\overline{L}_2 = 3, \overline{V}_2 = 21$ respectively. With the particulary choice of $\overline{\boldsymbol{\Lambda}}_2$, our framework is equivalent to macro-features of [3] whereas $\overline{\boldsymbol{\Lambda}}_1$ represents the classic SPM approach.

For all images converted in grey color, we extract $F = s_y \times s_x = 40 \times 40$ patches $\{O_k\}$ with size $m \times m = 26 \times 26$ pixels per scale. A total of 6400 patches are extracted per image for scales $\mathcal{S} = [1, 2, 3, 4]$. For sparse coding, we fixed $\beta = 0.2$. We trained models with a linear SVM *via* LIBLINEAR [6] ($C = 15$) with a *one-vs-all* multiclass approach. Accuracies are averaged over 10 folds cross-validation and best values of $K \in \{128, 256, 512, 1024, 2048\}$ found are noticed in the result's tables.

4.1 Scene-15 dataset

We use the Scene-15 dataset is containing a total of 4485 images assigned to $M = 15$ categories, with the number of images in each category ranging from 200 to 400. 100 images per class are used to train, the rest for testing. We obtained the second best result ($\mathbf{87.0\% \pm 0.48}$) without any sophisticated sparse coding such as LSc. For this latter, with SIFT patches, accuracy jumped from $80.28\% \pm 0.93$ to $89.75\% \pm 0.50$. We can expect such substantial gain with our approach using LSc. Notice also that KDES-EKM uses a concatenation of 3 descriptors coupled with an efficient features mapping (KDES-A+LSVM got $81.9\% \pm 0.60$ for a fair comparaison). We can also expect improvement using a specialized kernel during training such as χ^2 or HI kernels.

Algorithms	Accuracy ± Std
SIFT-ScSPM ($K = 1024, \underline{\boldsymbol{\Lambda}}_1, \overline{\boldsymbol{\Lambda}}_2$) [19]	$80.28\% \pm 0.93$
SIFT-MidLevel ($K = 2048, \underline{\boldsymbol{\Lambda}}_2, \overline{\boldsymbol{\Lambda}}_2$) [3]	$84.20\% \pm 0.30$
SIFT-LScSPM ($K = 1024, \underline{\boldsymbol{\Lambda}}_1, \overline{\boldsymbol{\Lambda}}_2$) [7]	$\mathbf{89.75\% \pm 0.50}$
KDES-EKM ($K = 1000, \underline{\boldsymbol{\Lambda}}_1, \overline{\boldsymbol{\Lambda}}_2$) [2]	*86.70%*
HC-ScSPM ($K = 2048, \underline{\boldsymbol{\Lambda}}_1, \overline{\boldsymbol{\Lambda}}_1$)	$86.05\% \pm 0.45$
HC-ScSPM ($K = 2048, \underline{\boldsymbol{\Lambda}}_2, \overline{\boldsymbol{\Lambda}}_1$)	$86.51\% \pm 0.52$
HIC-ScSPM ($K = 2048, \underline{\boldsymbol{\Lambda}}_1, \overline{\boldsymbol{\Lambda}}_1$)	$86.69\% \pm 0.44$
HIC-ScSPM ($K = 2048, \underline{\boldsymbol{\Lambda}}_2, \overline{\boldsymbol{\Lambda}}_1$)	$\mathbf{87.02\% \pm 0.48}$

Table 1. Classification rates on the Scene-15.

4.2 UIUC-Sport dataset

The UIUC-sport dataset is containing a total of 1579 images assigned to $M = 8$ categories. 60 images per class are used to train, 70 for testing. We outperform

Algorithms	Accuracy ± Std
SIFT-HOMP ($K = 2 \times 1024, \underline{\mathbf{\Lambda}}_1, \overline{\mathbf{\Lambda}}_2$) [1]	85.70%±1.30
SIFT-LScSPM ($K = 1024, \underline{\mathbf{\Lambda}}_1, \overline{\mathbf{\Lambda}}_2$) [7]	85.30%±0.31
SIFT-ScSPM ($K = 1024, \underline{\mathbf{\Lambda}}_1, \overline{\mathbf{\Lambda}}_2$) [19]	82.70%±1.50
HC-ScSPM ($K = 2048, \underline{\mathbf{\Lambda}}_1, \overline{\mathbf{\Lambda}}_1$)	86.56%±1.43
HIC-ScSPM ($K = 1024, \underline{\mathbf{\Lambda}}_1, \overline{\mathbf{\Lambda}}_1$)	$\mathbf{87.71\%\pm1.11}$

Table 2. Classification rates on the UIUC-Sport.

all previously published results on this dataset with $\mathbf{87.7\%\pm1.1}$ of accuracy even without using macro-features through $\underline{\mathbf{\Lambda}}_2$.

4.3 Caltech101 dataset

The Caltech101 dataset is containing a total of 9144 images assigned to $M = 102$ categories. 30 images per class are used to train, the rest for testing. In [15],

Algorithms	Accuracy ± Std
SIFT-LaRank ($K = 4096, \underline{\mathbf{\Lambda}}_1, \overline{\mathbf{\Lambda}}_2$) [15]	*80.02%±0.36*
SIFT-CDBN ($K = 4096, \underline{\mathbf{\Lambda}}_1, \overline{\mathbf{\Lambda}}_2$) [17]	77.80%±0.31
SIFT-multiway ($K = 1024, \underline{\mathbf{\Lambda}}_2, \overline{\mathbf{\Lambda}}_2$) [4]	77.30% ± 0.60
HC-ScSPM ($K = 1024, \underline{\mathbf{\Lambda}}_1, \overline{\mathbf{\Lambda}}_1$)	78.43%±0.27
HIC-ScSPM ($K = 1024, \underline{\mathbf{\Lambda}}_1, \overline{\mathbf{\Lambda}}_1$)	$\mathbf{79.05\%\pm0.33}$

Table 3. Classification rates on the Caltech101.

the *80.02%± 0.36* of accuracy with $K = 4096$ is obtained without indicating what kind of kernel they used in LaRank solver (probably a non-linear one). Our method with $\mathbf{79.05\%\pm0.33}$ of accuracy outperforms the Hierarchical CDBN approach of [17] even with a smaller dictionary size.

We have presented in this article the 2-layers BoS architecture mixing HLBP as local textures parametric encoder and Sc as non-parametric scenes encoder. Obtained performances outperform state-of-art results with a simple linear SVM. As potential future work, experimenting with LSc [7] and specialized kernels [2] will surely improve results.

References

[1] L. Bo, X. Ren, and D. Fox. Hierarchical matching pursuit for image classification: Architecture and fast algorithms. In *NIPS' 11*, pages 2115–2123.

[2] L. Bo, X. Ren, and D. Fox. Kernel descriptors for visual recognition. In *NIPS' 10*.

[3] Y. Boureau, F. Bach, Y. LeCun, and J. Ponce. Learning mid-level features for recognition. In *CVPR' 10*.

[4] Y. Boureau, N. Le Roux, F. Bach, J. Ponce, and Y. LeCun. Ask the locals: multi-way local pooling for image recognition. In *ICCV' 11*.

[5] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR' 05*.

[6] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. LIBLINEAR: A library for large linear classification. *JMLR*, 2008.

[7] S. Gao, I. W.-H. Tsang, L.-T. Chia, and P. Zhao. Local features are not lonely laplacian sparse coding for image classification. *Matrix*, 2010.

[8] M. Heikkilä, M. Pietikäinen, and C. Schmid. Description of interest regions with center-symmetric local binary patterns. In *CVGIP '06*.

[9] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR' 06*.

[10] S. Liao, X. Zhu, Z. Lei, L. Zhang, and S. Z. Li. Learning multi-scale block local binary patterns for face recognition. In *ICB*, 2007.

[11] D. G. Lowe. Object recognition from local scale-invariant features. In *ICCV' 99*.

[12] J. Mairal, F. Bach, J. Ponce, and G. Sapiro. Online dictionary learning for sparse coding. In *ICML '09*.

[13] T. Ojala, M. Pietikäinen, and T. Mäenpää. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Trans. PAMI*, 24(7), 2002.

[14] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, 42, 2001.

[15] G. L. Oliveira, E. R. Nascimento, A. W. Viera, and M. F. M. Campos. Sparse spatial coding: A novel approach for efficient and accurate object recognition. *ICRA' 12*.

[16] S. Paris and H. Glotin. Pyramidal multi-level features for the robot vision@icpr 2010 challenge. In *ICPR' 10*.

[17] K. Sohn, D. Y. Jung, H. Lee, and A. O. Hero III. Efficient Learning of Sparse , Distributed , Convolutional Feature Representations for Object Recognition. *ICCV' 11*.

[18] J. Wu and J. Rehg. Beyond the euclidean distance: Creating effective visual codebooks using the histogram intersection kernel. In *ICCV' 09*, pages 630 –637.

[19] J. Yang, K. Yu, Y. Gong, and T. S. Huang. Linear spatial pyramid matching using sparse coding for image classification. In *CVPR' 09*.

[20] B. Zhang, Y. Gao, S. Zhao, and J. Liu. Local derivative pattern versus local binary pattern: Face recognition with high-order local pattern descriptor. *IEEE Trans. Img. Proc.*, 19(2), 2010.

[21] L. Zhang, R. Chu, S. Xiang, S. Liao, and S. Z. Li. Face detection based on multi-block lbp representation. In *ICB' 07*.