Abstract—In this paper, we propose a three-layer spatial sparse coding (TSSC) for image classification, aiming at three objectives: naturally recognizing image categories without learning phase, naturally involving spatial configurations of images, and naturally counteracting the intra-class variances. The method begins by representing the test images in a spatial pyramid as the to-be-recovered signals, and taking all sampled image patches at multiple scales from the labeled images as the bases. Then, three sets of coefficients are involved into the cardinal sparse coding to get the TSSC, one to penalize spatial inconsistencies of the pyramid cells and the corresponding selected bases, one to guarantee the sparsity of selected images, and the other to guarantee the sparsity of selected categories. Finally, the test images are classified according to a simple image-to-category similarity defined on the coding coefficients. In experiments, we test our method on two publicly available datasets and achieve significantly more accurate results than the conventional sparse coding with only a modest increase in computational complexity.

Keywords—image classification; sparse coding; three-layer spatial sparse coding

I. INTRODUCTION

Nowadays, overwhelming amounts of images are available, which makes it imperative to design new classification algorithms capable of sifting through these images efficiently. This paper presents a method for image classification, which relaxes the learning phase of image classification tasks, softly involves spatial cues of natural images, and is robust to the intra-class variances.

The problem of image classification has been extensively studied and many breakthroughs have been made [1], [2], [3], [4]. The combination of bag-of-features (BoF) [5] and spatial pyramid matching (SPM) [1], [3] has recently dominated the scenario. Summarizing an image as a histogram of its local features, the BoF method is especially robust against spatial translations of features. However, totally disregarding spatial layout information of local features limits it to whole-image categorization tasks. This observation has triggered research efforts [1], [6] that take into account the spatial information for improving final performance, of which the combination with the SPM gained the most successful results[1], [3]. State-of-the-art as these methods are, they are built on the assumption that images belonging to a same category are roughly aligned, which is often violated in the real world. Furthermore, they pose a difficulty in scaling to thousands of categories as they require to learn a sophisticated model (usually by using SVM) for each category.

Another research stream in the literature prefers non-parametric classifiers with non-quantified image features [2], [7]. Boiman, etc. [2] represented both test images and labeled images by a group of image patches, and classified each test images by simply summarizing the category-recognizing results of its image patches. Undoubtedly, this method is more flexible than the SPM to account for large intro-class variabilities. However, it seems to go another extreme that the sampled image patches from an image are totally independent from each other, which still somewhat deviates from image nature.

In this paper, we propose a method for image classification based on a newly designed three-layer spatial sparse decomposition in ℓ1-norm sense, which seamlessly combines the merits of the two forementioned streams. We cast the image classification task to a sparse coding problem by representing the test images in a spatial pyramid as the to-be-recovered signals, and taking all sampled image patches at multiple scales from the labeled images as the bases. Importantly, we add a penalty term to the conventional sparse coding to penalize spatial inconsistencies between the pyramid cells and the corresponding selected bases, and introduce another two sparsities, image sparsity and category sparsity, in the traditional sparse coding to encourage the sparsity of selected images and selected categories. The underlying philosophy of our method is that all pyramid cells from an image can be more sparsely and consistently reconstructed via the patches (sparse I) from a few images (sparse II) belonging to the same category (sparse III). The two additional sparsities strive to enforce the selected bases into as few images and categories as possible, so striking a balance between visual similarities and belonging constraints. Images are classified by a simple image-to-category similarity defined on the sparse decomposition coefficients.

The remainder of this paper is organized as follows. We first formulate the TSSC in section II and demonstrate the classification method based on it in section III. Then, we devote section IV to experiments and analyses. Finally, we conclude this paper in section V.
II. THREE-LAYER SPATIAL SPARSE CODING

In this section, we first briefly review the principle of sparse coding and then show how we involve the spatial information of features into the sparse coding. Finally, we demonstrate the principle of the additional sparsity layers of the TSSC in detail.

A. Review of Sparse Coding

Sparse representation was put in a spotlight position in statistical signal processing community, especially after an exciting declaration that when the solution is sparse enough, it can be efficiently solved by convex $\ell_1$-norm minimization [8].

Suppose to solve a linear equation: $x = F\alpha$, where $x \in \mathbb{R}^m$ is the descriptor to be reconstructed, $\alpha \in \mathbb{R}^n$ is the vector of coding coefficients, and $F \in \mathbb{R}^{m \times n}$ is the $n$ bases in dictionary. When the assumption that $\alpha$ is enough sparse holds, it can be computed by the following convex $\ell_1$-norm optimization,

$$\arg\min_{\alpha} ||\alpha||_1, \text{s.t. } x = F\alpha. \quad (1)$$

B. Spatial Sparse Coding

In image representation, we adopt spatial pyramid for test images and use densely-sampled patches at multiple scales for labeled (‘training’) images. To make the descriptors of the sampled patches dimension-identical with that of the test images, we reorganize the descriptors of the sampled patches according to the employed pyramid. Specifically, if a pyramid with $pL$ levels is adopted, $(4^pL - 1)/3$ bases will be formed based on the descriptor of one sampled patch. These bases are formed by concatenating the extracted descriptor with $(4^pL - 1)/3 - 1$ length-identical zero vectors in all specifiable orders. For notational simplicity, we illustrate the reorganizations according to a two-layer spatial pyramid in Fig.1 (more-layer works in the same way). This representation mechanism is particularly suitable to sparse coding, as the test signal (spatial pyramid) is compact and informative and the bases (densely-sampled patches) are over-complete.

Furthermore, the bases are flexible, which is significant for the TSSC model. The reorganized bases have sharply different duties in recovering the test image. The base with the extracted descriptor in the first place can be, and only can be, selected to reconstruct signal of the first pyramid cell, the second to the second, ..., and the last to the last, as zero elements contribute zero to signal recovery. We refer to these cells as mother-cells to the corresponding bases, and thus, each base has a unique mother-cell to recover.

In order to introduce spatial cues, we add a spatial penalty term to the canonical sparse coding, referred to as spatial sparse coding (SSC). Pyramid cells are reconstructed by these reorganized bases under a soft spatial constraint, which encourages cells in the test images to be recovered by the bases at the ‘same’ location in the labeled images. Also, the multi-scale sampled bases provide additional scale invariants.

Let $D = \{I_1, I_2, \cdots, I_N\}$ denote the set of labeled images, the corresponding category labels are $C = \{c_1, c_2, \cdots, c_N\}$, where $c_i \in \{1, \cdots, L\}$, and $L$ is the number of categories. Then, we can embody the base dictionary $F$ as: $F = [f_{I_1,1}, \cdots, f_{I_1,N_1}, f_{I_2,1}, \cdots, f_{I_2,N_2}, \cdots, f_{I_N,1}, \cdots, f_{I_N,N_N}]$, where $K = (4^pL - 1)/3$, $F_{i,j}$ is the $j$th bases formed by the visual feature extracted from patch $j$ of image $i$. Let $a_{i,j}^k, \forall k \in (1, K)$, denote the normalized spatial distance (In our case, absolute Euclidean distance is used) between the $j$th patch of the $i$th image to its mother-cell, then, a set of variables, $\beta = [\beta_{i,j}^1, \cdots, \beta_{i,j}^K, \cdots, \beta_{i,n_1}^1, \cdots, \beta_{i,n_1}^K, \cdots, \beta_{n_2,n_2}^1, \cdots, \beta_{n_2,n_2}^K]^T$, are introduced to penalize spatial inconsistencies between the bases and their mother-cells,

$$\beta_{i,j}^k = \lambda_1 a_{i,j}^k \alpha_{i,j}. \quad (2)$$

By introducing a coefficient matrix $D \in \mathbb{R}^{O \times O}$, where $O = K \times \sum_{i=1}^{n_1} n_i$, and rearranging the index of $\alpha$ from 1 to $O$, according to the image indexes, the patch indexes in each image and the bases indexes of each patch, Eq.2 can be rewritten as:

$$\beta = \lambda_1 D\alpha \quad (3)$$
where $D$ is defined as follows:

$$
D = \begin{bmatrix}
  d_{1,1} & 0 & \cdots & 0 & \cdots & 0 \\
  0 & d_{1,1} & \cdots & 0 & \cdots & 0 \\
  \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\
  0 & 0 & \cdots & d_{N,n,N} & \cdots & 0 \\
\end{bmatrix}
$$

Thus, we need to optimize the following problem:

$$
\arg\min_{\alpha, \beta} \|\alpha\|_1 + \|\beta\|_1,
\text{s.t. } x = D\alpha, \quad \beta = \lambda_1 D\alpha
$$

(4)

The minimization of Eq.4 guarantees recovering sparsity and spatial consistency simultaneously, so softly introducing spatial cues when reconstructing the test images. We will reformulate it as a standard $\ell_1$-norm minimization in the next section.

C. Three-layer Spatial Sparse Coding

In a coding process, since the pyramid cells are extracted from one test image, it is natural to enforce the selected bases (patches) into a few individual images, which motivates the second layer of sparsity, referred to as image sparsity.

Let $S_{i,j}^k, \forall k \in (1, K)$, denote the normalized size of the $j$th patch of the $i$th image, we bring a set of coefficient $\gamma = [\gamma_1, \gamma_2, \ldots, \gamma_N]^T$ to measure the total weights of the labeled images,

$$
\gamma_i = \lambda_2 \sum_{j=1}^{N} \sum_{k=1}^{K} S_{i,j}^k \alpha_i^k
$$

(5)

By defining a matrix $S \in \mathbb{R}^{N \times O}$ using $S_{i,j}^k$,

$$
S = \begin{bmatrix}
  S_{1,1}^1 & \cdots & S_{1,n_1}^K & 0 & \cdots & 0 \\
  0 & \cdots & 0 & \cdots & 0 & \cdots & 0 \\
  \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
  0 & 0 & \cdots & S_{N,1}^1 & \cdots & S_{N,n,n}^K \\
\end{bmatrix}
$$

we can rewrite Eq.5 as

$$
\gamma = \lambda_2 S\alpha
$$

(6)

Then, we obtain the following optimization problem:

$$
\arg\min_{\alpha, \beta, \gamma} \|\alpha\|_1 + \|\beta\|_1 + \|\gamma\|_1,
\text{s.t. } x = F\alpha, \quad \beta = \lambda_1 D\alpha, \quad \gamma = \lambda_2 S\alpha
$$

(7)

Furthermore, since one image only belongs to one category, it is reasonable to enforce the selected images into a few categories, which motivates the third layer of sparsity, referred to as category sparsity. The category sparsity is defined below.

Similarly, we bring another set of coefficients, $\zeta = [\zeta_1, \zeta_2, \ldots, \zeta_L]^T$, to measure the weights of the categories,

$$
\zeta_l = \lambda_3 \sum_{i=1}^{N} \sum_{j=1}^{n_i} \sum_{k=1}^{K} \delta[c_i = l] S_{i,j}^k \alpha_i^k
$$

(8)

where $\delta[\cdot]$ is one if its argument is true and zero otherwise. Here, we define another matrix $B \in \mathbb{R}^{L \times O}$ using $S_{i,j}^k$ as:

$$
B = \begin{bmatrix}
  S_{1,1}^1 \delta_1^1 & \cdots & S_{1,n_1}^K \delta_1^l & \cdots & S_{N,n,n}^K \delta_1^N \\
  S_{1,1}^1 \delta_2^1 & \cdots & S_{1,n_1}^K \delta_2^l & \cdots & S_{N,n,n}^K \delta_2^N \\
  \vdots & \vdots & \vdots & \vdots & \vdots \\
  S_{1,1}^l \delta_L^1 & \cdots & S_{1,n_1}^K \delta_L^l & \cdots & S_{N,n,n}^K \delta_L^N \\
\end{bmatrix},
$$

where $\delta_l = \delta[c_i = l]$. Then Eq.8 can be rewritten as

$$
\zeta = \lambda_3 B\alpha
$$

(9)

Finally, we obtain the following optimization problem:

$$
\arg\min_{\alpha, \beta, \gamma, \zeta} \|\alpha\|_1 + \|\beta\|_1 + \|\gamma\|_1 + \|\zeta\|_1,
\text{s.t. } x = F\alpha, \quad \beta = \lambda_1 D\alpha, \quad \gamma = \lambda_2 S\alpha, \quad \zeta = \lambda_3 B\alpha
$$

(10)

Let

$$
\hat{x} = \begin{bmatrix}
  x \\
  0_{0 \times 1} \\
  0_{N \times 1} \\
  0_{L \times 1} \\
\end{bmatrix}, \quad \hat{\alpha} = \begin{bmatrix}
  \alpha \\
  \beta \\
  \gamma \\
  \zeta \\
\end{bmatrix},
$$

$$
\hat{F} = \begin{bmatrix}
  F, & 0_{m \times O}, & 0_{m \times N}, & 0_{m \times L} \\
  \lambda_1 D, & -I_{O \times O}, & 0_{O \times N}, & 0_{O \times L} \\
  \lambda_2 S, & 0_{N \times O}, & -I_{N \times N}, & 0_{N \times L} \\
  \lambda_3 B, & 0_{L \times O}, & 0_{L \times N}, & -I_{L \times L} \\
\end{bmatrix}
$$

(11)

where $m$ is the dimension of the visual feature. Then, we can reformulate the TSSC as a standard $\ell_1$-norm minimization problem as follows,

$$
\arg\min_{\hat{\alpha}} \|\hat{\alpha}\|_1, \text{s.t. } \hat{x} = \hat{F}\hat{\alpha}
$$

(12)

The minimization of Eq.11 guarantees the three above-mentioned sparsities simultaneously and introduces spatial constraints softly. Also, compared to Eq.1, Eq.11 increases only a modest computational complexity.

III. IMAGE CLASSIFICATION BY THE THREE-LAYER SPATIAL SPARSE CODING

Image classification is directly based on the coding coefficients $\hat{\alpha}$ obtained during the coding process. The probability of classifying image $I$ into category $l$ based on the TSSC is defined as:

$$
P_l(I) = \frac{\sum_{i=1}^{N} \sum_{j=1}^{n_i} \sum_{k=1}^{K} \max\{\hat{\alpha}_i^k \delta[c_i = l], 0\}}{\sum_{i'=1}^{N} \sum_{j'=1}^{n_{i'}} \sum_{k'=1}^{K} \max\{\hat{\alpha}_{i'}^{k'} \delta[c_{i'} = l'], 0\}}
$$

(12)

Many approaches [2], [4] to image classification have shown that the final classification results can be boosted by combining multiple types of features. In our case, when several (T) feature types are used, the decision rule linearly combines the image-to-category similarity of each feature (defined in Eq.12). Thus, the final classification rule is formulated as,

$$
\hat{l}(I) = \arg\max_{l} \sum_{i=1}^{T} \omega_l P_l^i(I)
$$

(13)
where $\omega_I$ is determined by the variance of the $t$th descriptor.

IV. EXPERIMENTS

In this section we test our method on the Caltech-101 category database [9], containing animals, furniture, vehicles, flowers, etc., and a 15 category image dataset, built gradually by [10], [6], and [1].

A. Implementation

We test our method with a combination of three descriptor types: the Gist descriptor [10], the SIFT descriptor [11], and the PACT descriptor [12]. Patches in labeled images are densely sampled at three scales: $64 \times 64$ pixels, $128 \times 128$ pixels, and $256 \times 256$ pixels, all with step of 32 pixels. We also compare our method to the nearest neighbor classifier (NN) with Euclidean distance (Eu) and histogram intersection (HI). For the NN classifier, training images are also represented by the spatial pyramid and the same feature combination mechanism, defined in Eq.13, is used. Gist descriptors and PACT descriptors are extracted directly from the sampled paths. SIFT descriptors are computed densely at patches of $16 \times 16$ pixels with step of 8 pixels, and then it is clustered into 200 clusters (words). Three different spatial pyramids ($pL = 1, 2, 3$) are tested in our experiments. We randomly select 30 and 100 images per category as labeled images (“training images”) for the Caltech-101 and the 15 category dataset, and report the average performance on 15 training-test partitions.

B. Results

The results on the two datasets are illustrated in Tab.1 and Tab.2, from which we can intuitively see that the two added sparsities, image sparsity and category sparsity, significantly boost classification performance. Another conclusion is that the spatial cues are really helpful in image classification.

Furthermore, the tables show us that our method significantly outperforms the NN classifier, which is can be attributed to two factors: (1) The soft spatial constraint used in the TSSC is superior to the rigid constraint in the NN classifier; (2) the TSSC strives to strike a balance between visual similarity, spacial consistency, and belonging constraint, which makes it more powerful in image classification.

V. CONCLUSION

This paper has presented a three-layer spatial sparse coding model for image classification. Our method, which involves three sets of coefficients into the traditional sparse coding to balance visual similarities, spacial configurations, and belonging constraints of natural images, has shown promising results on two publicly available datasets.

ACKNOWLEDGMENT

This work was supported by the grants from the National Natural Science Foundation of China (No. 40801183, 60890074) and by China 863 project 2008AA01Z126.