Tag Localization with Spatial Correlations and Joint Group Sparsity

Yang Yang Yi Yang Zi Huang Heng Tao Shen
The University of Queensland, Australia
{yang.yang,yi.yang,huang,shenht}@itee.uq.edu.au

Feiping Nie
University of Texas Arlington, USA
feipingnie@gmail.com

Abstract

Nowadays numerous social images have been emerging on the Web. How to precisely label these images is critical to image retrieval. However, traditional image-level tagging methods may become less effective because global image matching approaches can hardly cope with the diversity and arbitrariness of Web image content. This raises an urgent need for the fine-grained tagging schemes. In this work, we study how to establish mapping between tags and image regions, i.e. localize tags to image regions, so as to better depict and index the content of images. We propose the spatial group sparse coding (SGSC) by extending the robust encoding ability of group sparse coding with spatial correlations among training regions. We present spatial correlations in a two-dimensional image space and design group-specific spatial kernels to produce a more interpretable regularizer. Further we propose a joint version of the SGSC model which is able to simultaneously encode a group of intrinsically related regions within a test image. An effective algorithm is developed to optimize the objective function of the Joint SGSC. The tag localization task is conducted by propagating tags from sparsely selected groups of regions to the target regions according to the reconstruction coefficients. Extensive experiments on three public image datasets illustrate that our proposed models achieve great performance improvements over the state-of-the-art method in the tag localization task.

1. Introduction

Multimedia understanding and retrieval is a long standing research problem in the field of computer vision and multimedia [3, 12, 15, 21]. Nowadays, confronted with the huge number of social images on the Web, traditional image-level tagging methods tend to become less effective because global image matching approach can hardly handle the diversity and arbitrariness of Web image content. How to accurately tag images in more fine-grained levels becomes a great challenge to facilitate image retrieval. In this paper, we aim to address the problem of image tag localization, i.e., assigning tags to image regions.

Several related efforts have been made on this research topic. Multiple instance learning techniques [17, 18] and graph models [23] have been exploited and shown some effectiveness in region-level annotation. Most recently Liu *et al.* [5] proposed the Bi-Layer sparse coding for encoding image regions and propagating labels at region level. In this work, images were first segmented into basic regions, then the Bi-Layer model was applied to reconstruct each test region from a dictionary formed by other basic regions. The common tags of images containing the target region and sparsely selected regions will be re-assigned to the target region according to the reconstruction coefficients. It is worth noting that basic regions in the dictionary are implicitly assumed to be independent with each other. Contextual relationships among these semantic regions/objects, e.g., co-occurrence and spatial correlations, are ignored. Besides, when reconstructing regions within an image they individually encode each region and again ignore the intrinsic correlations among encoding regions. All of these correlations are important clues for uncovering the underlying data structure, and neglecting them may lead to a potential loss in interpretability and reconstruction performance. Hence, to overcome these drawbacks we propose a joint region reconstruction model which extends group sparse coding with collaborative encoding ability and integrates spatial correlations among basic regions into the training dictionary. Figure 1 illustrates our tag localization framework.

The contributions of this paper are summarized as follows: 1) We first propose the spatial group sparse coding (SGSC) which simultaneously takes advantage of the robust encoding ability of group sparse coding as well as prior knowledge about spatial correlations among image regions. We use the SGSC to encode an individual region from basic regions, thereby enabling tags to be propagated with the encoding coefficients; 2) Further, in order to collaboratively encode a group of regions in a test image, the Joint SGSC model is proposed by taking the intrinsic correlations among the group of test regions into consideration; 3) Moreover, a novel algorithm is developed to optimize the Joint

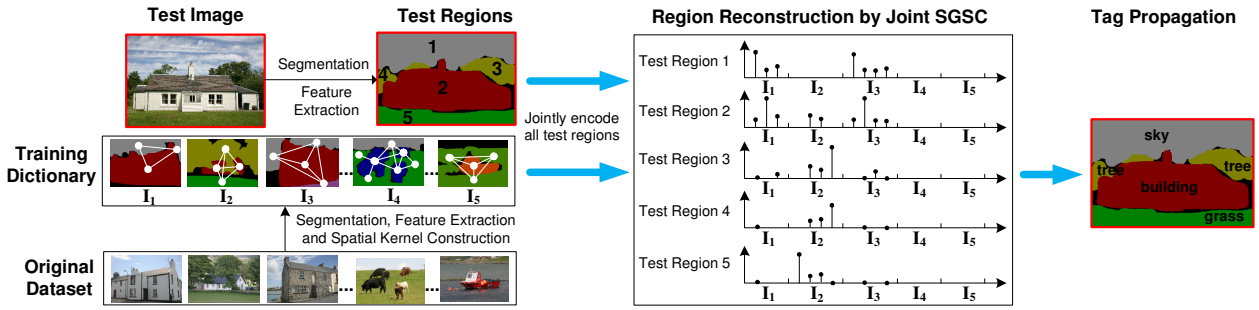

Figure 1. Overall illustration of tag localization framework with the Joint SGSC model. Given a test image, we first segment it and extract visual features for each region. All these test regions are simultaneously encoded from groups of spatially correlated regions (after segmentation, feature extraction and spatial kernel construction from the original training set). Finally, reconstruction coefficients are used to propagate tags from training regions to test regions.

SGSC. Theoretical proof and analysis are given to guarantee that the algorithm converges to the global optimality; 4) At last, extensive experiments are conducted on three public image datasets to show the effectiveness of our methods.

The rest of this paper is organized as follows. Related work will be reviewed in Section 2, followed by describing the details of the SGSC, the Joint SGSC and our proposed algorithm in Section 3. Section 4 reports all experimental results, followed by the conclusion in Section 5.

2. Related Work

The fundamental motivation of our work is to apply group sparse coding technique on the tag localization task by considering region correlations. We review research work on sparse coding and image tagging in this section.

In recent years, sparse coding has been a fairly popular technique in computer vision research. Yang *et al.* [19] improved vector quantization by extending sparse coding with spatial order of local descriptors. Gao *et al.* [3] followed this work and posed an additional constraint for enforcing maximal similarity preservation among similar descriptors. Mairal *et al.* [6] proposed simultaneous sparse coding to encode a group of similar patches for image restoration. Similarly in [1], Bengio *et al.* proposed a variant of sparse coding for jointly encoding the group of visual descriptors within the same image to achieve image-level sparsity. Wang *et al.* [14] proposed to use sparse coding twice for image annotation. They applied sparse coding to reconstruct images for establishing relations among images. The coefficients were used for dimensionality reduction over the feature representations. In [25], the authors applied group sparse coding to perform feature selection for image annotation. Our work is as well developed from sparse coding but different from Liu's work [5] mentioned in Section 1. Our model not only considers the intrinsic correlations amongst encoding regions but also explicitly integrates spatial correlations among basic regions to boost performance.

Image tag assignment is to automatically annotate an image with descriptive words. Wang *et al.* [15] collected candidate tags from surrounding textual information and re-ranked them based on visual information to acquire final tags. In [12], Siersdorfer *et al.* revealed the relationship among videos from the perspective of content redundancy, and proposed neighbor-based and context-based tagging schemes. In [20], Yang *et al.* handled the tag incompletion problem by grouping the visually near-duplicate images. Given a test image, a candidate tag set is first acquired from its near-duplicate neighbors. Then the candidate set is extended by using the multi-tag associations mined from the preprocessed image dataset. Finally, tag visual models are built for eliminating tag ambiguity. Most of these existing schemes perform tagging at image level whilst tag localization aims to assign tags to regions at more fine-grained levels.

3. Joint Spatial Group Sparse Coding

In this section, we propose a tag localization approach by uncovering how a group of regions can be jointly encoded from groups of spatially correlated basic regions.

3.1. Group Sparse Coding

Sparse representation has shown its effectiveness in computer vision due to the computational benefits and robustness. It assumes that a signal $y \in \mathbb{R}^d$ can be encoded by the sparse linear combination of N basic elements:

$$\hat{\beta}_{\ell_p} = \arg\min_{\beta} \frac{1}{2}\|y - X\beta\|_2^2 + \lambda\|\beta\|_p \qquad (1)$$

where $X \in \mathbb{R}^{d\times N}$ is the encoding dictionary, β indicates the encoding coefficients, λ is a trade-off parameter and $\|\cdot\|_p$ is the ℓ_p-norm. Ideally, the "pseudo-norm" ℓ_0-norm can guarantee to obtain the sparsest solution, but it has been proven to be an NP-hard selection problem. In practice one usually instead uses the ℓ_1-norm to reformulate sparse coding as a convex problem, which is known as the Lasso [13].

Although the Lasso enjoys significant computational strength and great performance, it is worth noting that this method implicitly assumes that an element in the dictionary is independent of all others. In image region encoding, if we simply concatenate regions of the training images to form the dictionary [5], apparently we lose correlation clues among regions within the same image, such as co-occurrence between objects (e.g., an image depicting a computer screen probably contains a computer keyboard), spatial dependency amongst regions (e.g., sky often lays over boats), etc. Besides, the Lasso tends to select more images because the ℓ_1-norm only guarantees region-level sparsity rather than image-level sparsity. This may introduce more potential noises when tag propagation is performed. Therefore, our first motivation is to integrate correlations among training regions and realize image-level sparsity.

Given a dictionary $X = \{X_1, X_2, \cdots, X_G\}$, where $X_g \in \mathbb{R}^{d \times N_g}$ consists of a group of N_g regions segmented from the g^{th} image, the Group Lasso [24] can be applied to reformulate the reconstruction process for a test region y:

$$\hat{\beta}_{gl} = \arg\min_{\beta} \frac{1}{2}\|y - \sum_{g=1}^{G} X_g\beta_g\|_2^2 + \lambda \sum_{g=1}^{G} \|\beta_g\|_2 \quad (2)$$

where $\beta_g \in \mathbb{R}^{N_g}$ is the encoding coefficient corresponding to the g^{th} group.

The Group Lasso uses a group-sparsity-inducing regularization instead of the ℓ_1-norm. In fact the regularization term $\lambda \sum_{g=1}^{G} \|\beta_g\|_2$ is the combination of both the ℓ_1-norm (inter-group) and the ℓ_2-norm (intra-group) and thus can be called the $\ell_1\ell_2$-norm. The fact that the Group Lasso considers multiple elements as a whole implies that it utilizes implicit relations among these elements to some extent. Nevertheless, in order to more precisely characterize the correlations we intend to explicitly integrate spatial correlations among groups of basic regions into the Group Lasso. Another restriction of the Group Lasso is that it can only encode one region at a time, which may lead to the loss of the intrinsic correlations and consistency among test regions.

3.2. Spatial Group Sparse Coding

As mentioned before neither the Lasso nor the Group Lasso explicitly considers correlations among basic elements. It has shown that such prior knowledge [16] is useful to uncover the characteristics of the data. In this subsection we focus on how to extend the Group Lasso with spatial correlations.

3.2.1 Spatial Dependency

Before reformulating our formula, we first handle how to represent spatial dependency among the semantic regions segmented from an image. We propose to describe the spatial relationships in a two-dimensional image space. In this space each region is represented by the 2-D coordinate of its center, and the Euclidean distance can be used to measure the spatial distance between two regions. To represent spatial dependency, we further propose to use Gaussian kernel to formulate the spatial similarities.

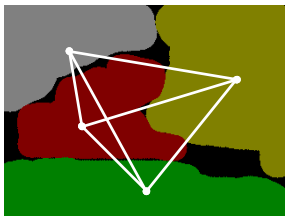

Figure 2. An illustration of an original image and the spatial dependency among its segmented regions. Each region is represented by the coordinate of its center and the edge between two regions is weighted by Gaussian similarity.

For a training image X_g and its N_g segmented regions, denote the two-dimensional coordinate of the i^{th} region as $\mathbf{z}_i^g \in \mathbb{R}^2$ in the image space. We build a Gaussian kernel matrix $K_g \in \mathbb{R}^{N_g \times N_g}$ over these regions' coordinates:

$$k_{ij}^g = k(\mathbf{z}_i^g, \mathbf{z}_j^g) = \exp\left(-\frac{\|\mathbf{z}_i^g - \mathbf{z}_j^g\|^2}{\gamma}\right) \quad (3)$$

where γ is the bandwidth parameter of Gaussian kernel.

Figure 2 illustrates an example of spatial dependency among regions, where each region is a vertex and the edge between two vertices is weighted by Gaussian similarity.

3.2.2 Extending Group Lasso with Spatial Prior

We propose the SGSC model by extending the Group Lasso with the spatial kernels mentioned before. In order to integrate K_g into the Group Lasso, we introduce a kernel norm based on β_g and K_g:

$$\|\beta_g\|_{K_g} = (\beta_g' K_g \beta_g)^{1/2} \quad (4)$$

We call this norm as the K_g-norm. If we substitute the ℓ_2-norm in Eq.(2) with the K_g-norm, our SGSC model is obtained as follows:

$$\hat{\beta}_{sgl} = \arg\min_{\beta} \frac{1}{2}\|y - \sum_{g=1}^{G} X_g\beta_g\|_2^2 + \lambda \sum_{g=1}^{G} \|\beta_g\|_{K_g} \quad (5)$$

In [24], it has shown that the kernel group regularization used in Eq.(5) is an intermediate regularizer between the ℓ_1-norm in the Lasso and the ℓ_2-norm in ridge regression. This means that our SGSC model encourages image-level sparsity when taking spatial dependency prior knowledge into consideration. Similar to the Group Lasso, the SGSC model does not consider the intrinsic correlations among test regions either, thereby leading to potential loss in the reconstruction performance and consistency. We will address this limitation in the following subsection.

3.3. Region Reconstruction by Joint SGSC

In this subsection we aim at addressing the limitation of the SGSC model mentioned before, i.e., the intrinsic correlations of test regions are ignored. We propose a joint version of the SGSC to simultaneously encode all test regions within the same image.

3.3.1 Joint SGSC

Considering a group of test regions segmented from an image, surely we can reconstruct each region individually to fulfill the tag localization task for the test image. However, even though the SGSC provides image-level sparsity for the individual region encoding task, it cannot guarantee consistency and robustness in encoding all test regions of the whole image since they are intrinsically correlated with each other. Therefore, we formulate image reconstruction as a joint region reconstruction procedure and propose the Joint SGSC model. The Joint SGSC helps achieve that once a group of training regions have been chosen as sparse codes for one test region, then probably should they be chosen to represent other test regions within the same image without adding much extra penalty cost.

Denote $Y \in \mathbb{R}^{d\times N_x}$ as N_x test regions segmented from a test image x. In previous works [1][6], the authors proposed a simultaneous sparsity regularizer for jointly reconstructing Y as follows:

$$\mathcal{B}^* = \arg\min_{\mathcal{B}} \frac{1}{2}\|Y - X\mathcal{B}\|_F^2 + \lambda \sum_{i=1}^{N} \|\beta_i\|_p \qquad (6)$$

where the first term penalizes the whole reconstruction error of all encoding regions. $\mathcal{B} = [\beta_1^T, \beta_2^T, \ldots, \beta_N^T]^T \in \mathbb{R}^{N\times N_x}$ denotes the reconstruction coefficient matrix, and $\beta_i \in \mathbb{R}^{1\times N_x}$ (the i^{th} row of matrix $\mathcal{B}$) specifies the contribution of the i^{th} basic region of X to each test region. In essence, (6) is the joint version of the Lasso. The regularizer $\lambda\sum_{i=1}^{N}\|\beta_i\|_p$ tends to minimize the number of nonzero rows in $\mathcal{B}$. As with the Lasso, this formulation ignores correlations among basic elements and only guarantees element-level sparsity rather than group-level.

To overcome these drawbacks we extend the K_g-norm to a joint version which guarantees sparsity when jointly reconstructing all test regions:

$$\|\mathcal{B}_g\|_{K_g} = (\sum_{j=1}^{N_x} \|\beta_j^g\|_{K_g}^2)^{1/2} = (tr(\mathcal{B}_g^T K_g \mathcal{B}_g))^{1/2} \qquad (7)$$

where $tr(\cdot)$ is the trace of a matrix. $\mathcal{B} = (\mathcal{B}_1^T, \mathcal{B}_2^T, \ldots, \mathcal{B}_G^T)^T$ and each $\mathcal{B}_g \in \mathbb{R}^{N_g\times N_x}$ specifies the contribution of the g^{th} group to all encoding regions. β_j^g is the j^{th} column of $\mathcal{B}_g$ and $\|\beta_j^g\|_{K_g}$ is the K_g-norm on β_j^g as defined in Eq.(4). Hereby, we define a new spatial kernel regularizer by summing over all training groups and the Joint SGSC is finally proposed as follows:

$$\mathcal{B}^* = \arg\min_{\mathcal{B}} \frac{1}{2}\|Y - X\mathcal{B}\|_F^2 + \lambda\|\mathcal{B}\|_K \qquad (8)$$

where $\|\mathcal{B}\|_K = \sum_{g=1}^{G}\|\mathcal{B}_g\|_{K_g}$. As we can see, our formula is the natural extension of the SGSC model and (6). In the next section we develop an effective iterative algorithm to optimize the objective function of (8).

3.3.2 Computation of Joint SGSC

In order to optimize (8) we first transform the K_g-norm into the *Frobenius* norm and then propose an effective algorithm to find the global optimality.

For a training image X_g, without loss of generality we assume all its regions have different coordinates, then the spatial gaussian kernel matrix K_g derived from these coordinates is symmetric positive-definite [7]. We perform Cholesky decomposition on the kernel matrix: $K_g = U_g^T U_g$. Here U_g is an upper triangular matrix with strictly positive diagonal entries. It is clear that U_g is invertible. After substituting the above result into Eq. (7) we obtain:

$$\|\mathcal{B}_g\|_{K_g} = tr(\mathcal{B}_g^T U_g^T U_g \mathcal{B}_g)^{1/2} = tr(\mathcal{A}_g^T \mathcal{A}_g)^{1/2} = \|\mathcal{A}_g\|_F$$

where $\mathcal{A}_g = U_g\mathcal{B}_g$ and $\|\cdot\|_F$ is the *Frobenius* norm. Then the objective function of (8) can be rewritten as below:

$$\begin{aligned} &\min_{\mathcal{B}} \frac{1}{2}\|Y - \sum_{g=1}^{G} X_g\mathcal{B}_g\|_F^2 + \lambda\sum_{g=1}^{G}\|\mathcal{B}_g\|_{K_g} \\ \Leftrightarrow \quad &\min_{\mathcal{A}} \frac{1}{2}\|Y - \sum_{g=1}^{G} \tilde{X}_g\mathcal{A}_g\|_F^2 + \lambda\sum_{g=1}^{G}\|\mathcal{A}_g\|_F \end{aligned} \qquad (9)$$

where $\tilde{X}_g = X_g U_g^{-1}, g = 1, 2, \ldots, G$.

Further, we propose an effective iterative algorithm (as illustrated in Algorithm 1) to optimize (9). In contrast to coordinate descent based algorithms, in each iteration our algorithm directly obtains an analytical solution which guarantees the decreasing trend of (9). We will show that this algorithm guarantees that $\mathcal{A}$ converges to the globally optimal solution. In [8], an iterative algorithm was proposed to solve the joint $\ell_{2,1}$-norm minimization problem. Inspired by [8], we derive the Theorem 1 and prove that Algorithm 1 can obtain the globally optimal solution for (9). Before that we introduce two lemmas.

Lemma 1. *Denote $\mathcal{A}$ as the optimal result of the t^{th} iteration and $\tilde{\mathcal{A}}$ as the variable of $(t+1)^{th}$ iteration of Algorithm 1, then the following inequality holds:*

$$\left\|\tilde{\mathcal{A}}\right\|_F - \frac{tr(\tilde{\mathcal{A}}^T\tilde{\mathcal{A}})}{2\left\|\mathcal{A}\right\|_F} \leqslant \left\|\mathcal{A}\right\|_F - \frac{tr(\mathcal{A}^T\mathcal{A})}{2\left\|\mathcal{A}\right\|_F}$$

Algorithm 1: An effective iterative algorithm for optimizing the Joint SGSC model.

Input : Original data matrix X, spatial kernels $K_g(g = 1, 2, \cdots, G)$, observation data matrix Y and initialized coefficient $\mathcal{B}$.

Output: Globally optimal encoding coefficients $\mathcal{B}^*$.

```
for g = 1 to G do
    Perform Cholesky decomposition: K_g = U_g^T U_g;
    X̃_g = X_g U_g^{-1};
    A_g = U_g B_g;
repeat
    Let D = diag(||A_1||_F I_1, ..., ||A_G||_F I_G);
    A = (D X̃^T X̃ + λI)^{-1} D X̃^T Y;
until there is no change to A;
for g = 1 to G do
    B*_g ← U_g^{-1} A_g;
```

Proof. See Appendix. □

Lemma 2. *Given $\mathcal{A} = [\mathcal{A}_1^T \mathcal{A}_2^T \cdots \mathcal{A}_G^T]^T$, $\mathcal{A}_g$ is a submatrix of $\mathcal{A}$ and corresponds to the g^{th} group, then we have the following conclusion:*

$$\sum_{g=1}^{G} \left\|\tilde{\mathcal{A}}_g\right\|_F - \sum_{g=1}^{G} \frac{tr(\tilde{\mathcal{A}}_g{}^T \tilde{\mathcal{A}}_g)}{2\left\|\mathcal{A}_g\right\|_F} \leqslant \sum_{g=1}^{G} \left\|\mathcal{A}_g\right\|_F - \sum_{g=1}^{G} \frac{tr(\mathcal{A}_g{}^T \mathcal{A}_g)}{2\left\|\mathcal{A}_g\right\|_F}$$

Proof. See Appendix. □

Now we come out with the conclusion of Theorem 1 according to Lemma 1 and 2.

Thereom 1. *At each iteration of Algorithm 1, the value of the objective function in (9) monotonically decreases.*

Proof. We first optimize the following quadratic problem:

$$\min_{\tilde{\mathcal{A}}} \frac{1}{2}\left\|Y - \tilde{X}\tilde{\mathcal{A}}\right\|_F^2 + \lambda \sum_{g=1}^{G} \frac{tr(\tilde{\mathcal{A}}_g{}^T \tilde{\mathcal{A}}_g)}{2\left\|\mathcal{A}_g\right\|_F} \qquad (10)$$

By setting the deviation of (10) w.r.t. $\tilde{\mathcal{A}}$ to zero we obtain $\tilde{\mathcal{A}}^* = (D\tilde{X}^T\tilde{X} + \lambda I)^{-1} D\tilde{X}^T Y$, where

$$D = \begin{bmatrix} \left\|\mathcal{A}_1\right\|_F I_1 & & \\ & \ldots & \\ & & \left\|\mathcal{A}_G\right\|_F I_G \end{bmatrix}$$

is a diagonal matrix orderly formed by G sub diagonal matrices corresponding to G groups. Then we respectively substitute $\tilde{\mathcal{A}}^*$ and $\mathcal{A}$ into (10) and get:

$$\begin{aligned}
&\frac{1}{2}\left\|Y - \tilde{X}\tilde{\mathcal{A}}^*\right\|_F^2 + \lambda \sum_{g=1}^{G} \frac{tr(\tilde{\mathcal{A}}_g^{*T} \tilde{\mathcal{A}}_g{}^*)}{2\left\|\mathcal{A}_g\right\|_F} \leqslant \frac{1}{2}\left\|Y - \tilde{X}\mathcal{A}\right\|_F^2 \\
&\qquad + \lambda \sum_{g=1}^{G} \frac{tr(\mathcal{A}_g{}^T \mathcal{A}_g)}{2\left\|\mathcal{A}_g\right\|_F} \\
\Rightarrow\ &\frac{1}{2}\left\|Y - \tilde{X}\tilde{\mathcal{A}}^*\right\|_F^2 + \lambda \sum_{g=1}^{G} \left\|\tilde{\mathcal{A}}_g^*\right\|_F - \lambda \sum_{g=1}^{G} \left(\left\|\tilde{\mathcal{A}}_g^*\right\|_F \right. \\
&\qquad \left. - \frac{tr(\tilde{\mathcal{A}}_g^{*T} \tilde{\mathcal{A}}^*)}{2\left\|\mathcal{A}_g\right\|_F}\right) \leqslant \frac{1}{2}\left\|Y - \tilde{X}\mathcal{A}\right\|_F^2 + \lambda \sum_{g=1}^{G} \left\|\mathcal{A}_g\right\|_F \\
&\qquad - \lambda \sum_{g=1}^{G} \left(\left\|\mathcal{A}_g\right\|_F - \frac{tr(\mathcal{A}_g{}^T \mathcal{A}_g)}{2\left\|\mathcal{A}_g\right\|_F}\right) \\
\Rightarrow\ &\frac{1}{2}\left\|Y - \tilde{X}\tilde{\mathcal{A}}^*\right\|_F^2 + \lambda \sum_{g=1}^{G} \left\|\tilde{\mathcal{A}}_g^*\right\|_F \leqslant \frac{1}{2}\left\|Y - \tilde{X}\mathcal{A}\right\|_F^2 \\
&\qquad + \lambda \sum_{g=1}^{G} \left\|\mathcal{A}_g\right\|_F
\end{aligned}$$

As we can see that at $(t+1)^{th}$ iteration $\tilde{\mathcal{A}}^*$ indeed makes the value of Eq. (9) decreased. □

Because of the convexity of objective function (9), Theorem 1 clearly guarantees that Algorithm 1 converges to the global optimality. After obtaining the encoding coefficients of a test region, we determine its tag as follows. For each tag, we accumulate the coefficients of the training regions that are associated with this tag. Then the tag with the highest score is chosen as the final tag for the test region.

4. Experiments

In this section we employ three public image datasets to evaluate the effectiveness of our proposed SGSC and Joint SGSC models on the tag localization task.

4.1. Experimental Settings

4.1.1 Datasets

In our work three image datasets (two versions of MSRC [11] and SAIAPR TC-12 [2]) with region-level ground truths are used to evaluate our proposed models. The pixel-wise labeled MSRC dataset contains two versions: Version 1 provides 240 images and 13 labels while Version 2 is comprised of 591 images and 23 labels. Both of them have been manually segmented and labeled at pixel level. The SAIAPR TC-12 contains about 20,000 images and it also provides: 1) Segmentation masks and segmented images; 2) Region-level Features and Labels. According to our observation, the SAIAPR is organized into 40 subsets and each subset contains relatively relevant images (e.g. images taken at the same landscape). We choose the one containing

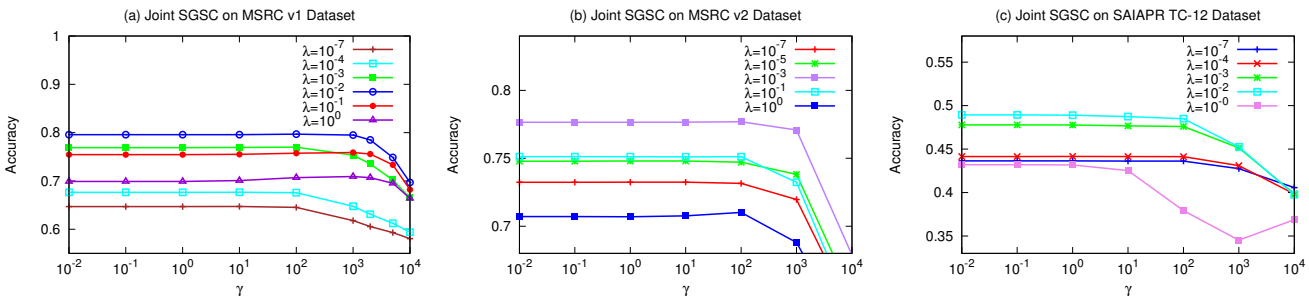

Figure 3. Effects of parameters λ in sparse coding formula and γ in spatial kernel for accuracy of the Joint SGSC on 3 datasets.

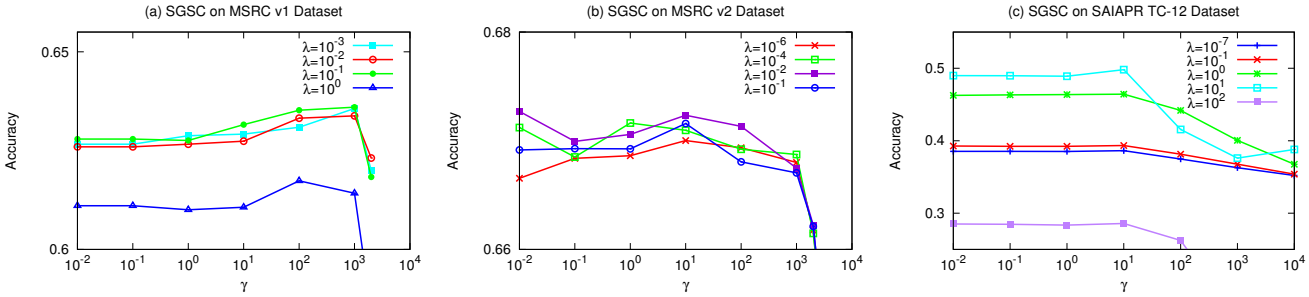

Figure 4. Effects of parameters λ in sparse coding formula and γ in spatial kernel for accuracy of the SGSC on 3 datasets.

251 images. Note that our tagging method can be easily extended to large-scale datasets by preliminarily filtering out those potentially relevant images.

4.1.2 Images Segmentation and Features Extraction

The MSRC datasets do not provide segmentation masks and region-level features so we preliminarily segment the images and extract visual features from segmented regions. Various image segmentation algorithms [10, 22] can be used here and we choose to use Normalized Cuts Clustering [10] in this paper. Images in the MSRC have 3-4 labels in average. So we set the number of segmentations to 4. We cannot expect Normalized Cuts to generate the same segmentation results as manual ground truths. Therefore, we assign each region with its dominant label. For both MSRC and SAIAPR datasets, we characterize visual content of their image regions by extracting Local Binary Patterns (LBP) feature [9]. LBP assigns each pixel with a value by comparing its 8 neighbor pixels with the center pixel value and transforming the result to a binary value. Then the histogram of the values is accumulated as a local descriptor.

4.1.3 Comparing Algorithms

We first choose the Lasso and the Group Lasso as two basic baselines to show the effectiveness of our consideration on spatial correlations among basic training regions. We also implement the SGSC as another baseline in that we want to illustrate the advantage of the joint encoding ability of the Joint SGSC. Moreover we intend to compare with the Bi-Layer sparse coding [5] which is one of the state-of-the-art algorithms for the region-level tagging task. To keep consistence with the settings of baselines and our algorithm, we slightly modify the formula of the Bi-Layer sparse coding to an extended Lasso. For the Lasso, the Group Lasso and the Bi-Layer sparse coding we use the implementation of SLEP package [4]. At last, to illustrate the general performance we also compare our models with the classical kNN algorithm and k is empirically set to 50 and 100.

4.2. Parameter Setting

We test different parameter settings for our models to obtain the best experimental results. There are two parameters: 1) λ is used to keep balance between reconstruction error and the level of sparsity of the encoding coefficients. By default we set $\lambda \in \{10^{-7}, 10^{-6}, \cdots, 10^{2}\}$; 2) γ used in Gaussian spatial kernel tunes the effect of region spatial dependency prior. Here we set $\gamma \in \{10^{-2}, 10^{-1}, \cdots, 10^{4}\}$ in the following experiments.

The results on average tagging accuracy for the Joint SGSC and the SGSC on three datasets are illustrated in Figure 3 and 4 respectively. In order to clearly characterize the effects of λ and γ, we report the partial experimental results for brevity. It is clear that the Joint SGSC outperforms the SGSC in most cases. Let us first see the effect of λ (inter-curve comparison). Both Figure 3 and 4 show that as λ increases the average accuracy does not monotonically increase. Best λ (top curve in each sub-figure) all appear in the middle of the range, which means both the reconstruc-

Datasets	kNN ($k = 50$)	kNN ($k = 100$)	Lasso	Group Lasso	Bi-Layer	SGSC	Joint SGSC
MSRC_v1	0.364	0.315	0.630	0.630	0.632	0.640	**0.797**
MSRC_v2	0.448	0.434	0.673	0.671	0.674	0.672	**0.777**
SAIAPR TC-12	0.185	0.174	0.384	0.490	0.385	**0.498**	0.489

Table 1. Overall average accuracy comparison of different algorithms on MSRC and SAIAPR TC-12 datasets. We use k=50 and 100 for kNN method and the best results for the Lasso, the Group Lasso, the Bi-Layer sparse coding and our proposed SGSC and Joint SGSC.

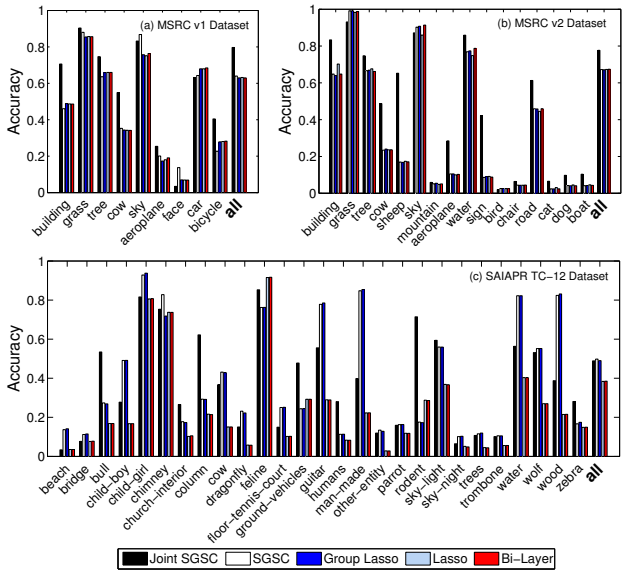

Figure 5. Detailed tag localization performance.

tion error term and the jointly spatial sparsity regularizer play important roles in finding the optimal region encoding results. We observe the similar results when testing the effect of γ (intra-curve comparison). Most lines go up slowly as γ increases from a very small value, then drop from their peak accuracies as γ becomes very large. Taking Figure 4 (a) as an example, it shows a clear trend. Actually, when γ is extremely large, spatial transformation becomes overwhelming, which may lead to some negative distortion to the reconstruction performance. On the other hand when γ is small, spatial kernels are close to the identity matrix, which ignores spatial transformation. Similar trends can be seen in other sub-figures of Figure 3 and 4.

In summary the average accuracy of our models could achieve the best results by testing different combinations of λ and γ. For fair comparison we also respectively tune parameters for the Lasso, the Group Lasso, and the Bi-Layer sparse coding to find their best performance on different datasets. In the following experiments, for all algorithms the best results will be used for comparison.

4.3. Comparison and Analysis

In this subsection we report and analyze the results of our proposed models and comparing algorithms in terms of the average tag localization accuracy.

Table 1 illustrates the overall average tag localization accuracies of different algorithms on different datasets. Detailed tag localization performance comparisons on each tag are reported in Figure 5. As we can see from Table 1, on both MSRC_v1 and MSRC_v2 datasets, the performance of the Lasso, the Group Lasso, the Bi-Layer sparse coding and the SGSC is very close. However, the Joint SGSC consistently outperforms all of them by nearly 20%, which is huge. This phenomenon apparently indicates the effectiveness of integrating the spatial dependency and joint encoding ability in our joint model. We believe that the Joint SGSC truly preserves the consistency of the intrinsic correlations among the test regions, and the spatial dependency among the training regions also poses positive effect on the reconstruction performance. As to the SAIAPR TC-12 dataset, we find that both the SGSC and the Joint SGSC obtain more than 10% improvement over the Lasso and the Bi-Layer sparse coding. Different from the MSRC datasets, the performance of the SGSC on the SAIAPR dataset slightly beats the Joint SGSC. The reason why joint encoding degrades is that images in the SAIAPR dataset are more arbitrary, hence the regions of these images do not have inevitable relationships and the joint model cannot effectively reveal the real correlations among encoding regions. From the detailed tagging performance in Figure 5 we find that our models usually obtain better performance on category of animal, such as *cow*, *rodent*, *zebra*, etc. Normally animals do not appear alone in the dataset, i.e., there are usually more than one animals in an image. This gives a strong hint that the considered joint relationship among test regions is effective in our Joint SGSC. In contrast, we do not achieve better results on tags that often correspond to one single object in an image, such as *sky*, etc.

Note that in our experiments we obtain different results on the MSRC_v2 dataset for the Bi-Layer sparse coding and the Lasso from those reported in [5]. We think the reasons exist in several aspects: 1) Different preprocessing strategies lead to different subsets of the MSRC_v2 dataset, which may affect the performance; 2) The slight modification to the formula probably causes the results sensitive to the parameters; 3) The segmentation methods adopted in these two works are not the same, which may create different ground truths of segmented regions, thereby leading to different performances; 4) Most importantly, tags are propagated in different levels, i.e., they propagated tags from

image to region while we assign tags from region to region.

5. Conclusion and Future Work

In this paper we develop the SGSC and the Joint SGSC models for localizing tags to image regions. Besides integrating spatial relationships among training regions, we also realize the joint encoding of a group of intrinsically relevant regions within the test image. An effective algorithm is further proposed to optimize the Joint SGSC, and we also discuss and prove its convergence to the global optimality. Finally we conduct extensive experiments on three public image datasets to show the superiority of our proposals. In future we intend to apply kernel methods to adapt our model to nonlinear settings.

A. Proof of Lemma 1 and 2

We first prove Lemma 1 and then Lemma 2.

Proof. Since $\|\tilde{\mathcal{A}}\|_F$ and $\|\mathcal{A}\|_F$ are real values, we have:

$$\begin{aligned}
&(\|\tilde{\mathcal{A}}\|_F - \|\mathcal{A}\|_F)^2 \geqslant 0 \\
\Rightarrow &tr(\tilde{\mathcal{A}}^T\tilde{\mathcal{A}}) + tr(\mathcal{A}^T\mathcal{A}) - 2\|\tilde{\mathcal{A}}\|_F \|\mathcal{A}\|_F \geqslant 0 \\
\Rightarrow &2\|\tilde{\mathcal{A}}\|_F \|\mathcal{A}\|_F - tr(\tilde{\mathcal{A}}^T\tilde{\mathcal{A}}) \leqslant tr(\mathcal{A}^T\mathcal{A}) \\
\Rightarrow &\|\tilde{\mathcal{A}}\|_F - \frac{tr(\tilde{\mathcal{A}}^T\tilde{\mathcal{A}})}{2\|\mathcal{A}\|_F} \leqslant \frac{tr(\mathcal{A}^T\mathcal{A})}{2\|\mathcal{A}\|_F} \\
\Rightarrow &\|\tilde{\mathcal{A}}\|_F - \frac{tr(\tilde{\mathcal{A}}^T\tilde{\mathcal{A}})}{2\|\mathcal{A}\|_F} \leqslant \|\mathcal{A}\|_F - \frac{tr(\mathcal{A}^T\mathcal{A})}{2\|\mathcal{A}\|_F}
\end{aligned}$$

□

Proof. According to Lemma 1, for each group g we have:

$$\left\|\tilde{\mathcal{A}}_g\right\|_F - \frac{tr(\tilde{\mathcal{A}}_g^T\tilde{\mathcal{A}}_g)}{2\|\mathcal{A}_g\|_F} \leqslant \|\mathcal{A}_g\|_F - \frac{tr(\mathcal{A}_g^T\mathcal{A}_g)}{2\|\mathcal{A}_g\|_F}$$

Thus, by summing over all G above inequalities we get the conclusion of Lemma 2. □

References

[1] S. Bengio, F. Pereira, Y. Singer, and D. Strelow. Group sparse coding. In *NIPS*, pages 90–98, 2009.

[2] H. J. Escalante, C. A. Hernandez, J. A. Gonzalez, and A. Loez-Loez. The segmented and annotated iapr tc-12 benchmark. *CVIU*, 114:419–428, 2010.

[3] 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*, pages 3555–3561, 2010.

[4] J. Liu, S. Ji, and J. Ye. *SLEP: Sparse Learning with Efficient Projections*. Arizona State University, 2009.

[5] X. Liu, B. Cheng, S. Yan, J. Tang, and T. S. Chua. Label to region by bi-layer sparsity priors. In *ACM Multimedia*, pages 115–124, 2009.

[6] J. Mairal, F. B. Jean, Ponce, G. Sapiro, and A. Zisserman. Non-local sparse models for image restoration. In *ICCV*, pages 2272–2279, 2009.

[7] C. A. Micchelli. Interpolation of scattered data: Distance matrices and conditionally positive definite functions. *Constructive Approximation*, 2:11–22, 1986.

[8] F. Nie, H. Huang, X. Cai, and C. Ding. Efficient and robust feature selection via joint ℓ2,1-norms minimization. In *NIPS*, 2010.

[9] T. Ojala, M. Pietikainen, and D. Harwood. A comparative study of texture measures with classification based on featured distributions. *PR*, 29:51–59, 1996.

[10] J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE TPAMI*, 22:888–905, 2000.

[11] J. Shotton, J. Winn, C. Rother, and A. Criminisi. Textonboost for image understanding: Multi-class object recognition and segmentation by jointly modeling texture, layout, and context. *IJCV*, 81:2–23, 2009.

[12] S. Siersdorfer, J. San Pedro, and M. Sanderson. Automatic video tagging using content redundancy. In *SIGIR*, pages 395–402, 2009.

[13] R. Tibshirani. Regression shrinkage and selection via the lasso. *JRSSB*, 58:267–288, 1996.

[14] C. Wang, S. Yan, L. Zhang, and H.-J. Zhang. Multi-label sparse coding for automatic image annotation. In *CVPR*, pages 1643–1650, 2009.

[15] X.-J. Wang, L. Zhang, X. Li, and W.-Y. Ma. Annotating images by mining image search results. *IEEE TPAMI*, 30:1919–1932, 2008.

[16] Z. J. Xiang, Y. taylor Xi, U. Hasson, and P. J.Ramadge. Boosting with spatial regularization. In *NIPS*, pages 2107–2115, 2009.

[17] C. Yang, M. Dong, and F. Fotouhi. Region based image annotation through multiple-instance learning. In *ACM Multimedia*, pages 435–438, 2005.

[18] C. Yang, M. Dong, and J. Hua. Region-based image annotation using asymmetrical support vector machine-based multiple-instance learning. *CVPR*, pages 2057–2063, 2006.

[19] J. Yang, K. Yu, Y. Gong, and T. Huang. Linear spatial pyramid matching using sparse coding for image classification. In *CVPR*, pages 1794–1801, 2009.

[20] Y. Yang, Z. Huang, H. Shen, and X. Zhou. Mining multi-tag association for image tagging. *WWW*, pages 1–24, 2010.

[21] Y. Yang, D. Xu, F. Nie, J. Luo, and Y. Zhuang. Ranking with local regression and global alignment for cross media retrieval. In *ACM Multimedia*, pages 175–184, 2009.

[22] Y. Yang, D. Xu, F. Nie, S. Yan, and Y. Zhuang. Image clustering using local discriminant models and global integration. *IEEE TIP*, 19:2761–2773, 2010.

[23] J. Yuan, J. Li, and B. Zhang. Exploiting spatial context constraints for automatic image region annotation. In *ACM Multimedia*, pages 595–604, 2007.

[24] M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *JRSSB*, 68:49–67, 2006.

[25] S. Zhang, J. Huang, Y. Huang, Y. Yu, H. Li, and D. N. Metaxas. Automatic image annotation using group sparsity. In *CVPR*, pages 3312–3319, 2010.