Learning Sparse Covariance Patterns for Natural Scenes

Liwei Wang[†] Yin Li[‡] Jiaya Jia[†] Jian Sun[§] David Wipf[§] James M. Rehg[‡]

[†]The Chinese University of Hong Kong [‡]Georgia Institute of Technology [§]Microsoft Research Asia

Abstract

For scene classification, patch-level linear features do not always work as well as handcrafted features. In this paper, we present a new model to greatly improve the usefulness of linear features in classification by introducing covariance patterns. We analyze their properties, discuss the fundamental importance, and present a generative model to properly utilize them. With this set of covariance information, in our framework, even the most naive linear features that originally lack the vital ability in classification become powerful. Experiments show that the performance of our new covariance model based on linear features is comparable with or even better than handcrafted features in scene classification.

1. Introduction

Finding appropriate feature representation for visual data, i.e., images and videos, is central to computer vision tasks due to its high importance in solving many recognition and classification problems. Existing visual features can be approximately classified into two categories, i.e., handcrafted features and features learned automatically from image data. Both of them have been extensively employed and evaluated in different applications and have exhibited different properties.

Specifically, handcrafted features, such as SIFT [17] and HoG [4], are manually designed based on histograms of dominant gradient orientation in local regions. CENTRIST [33] summarizes local shape and texture information via histogram of Census Transform. When fit into spatial pyramid matching (SPM) [14], these features achieve state-of-the-art results [34, 36].

In the meantime, learning based features [2, 23, 10, 38] mainly follow the generative interpretation that is closely related to human perceptual understanding of natural images. Sparse coding is a popular representative within this class. Originating from the efficient coding theory that explains human early vision system [30], sparse representation is commonly modeled as a linear combination of basis vectors over a pre-learned over-complete dictionary and has been widely used in low-level vision [35, 5]. However, when applied to classification, it seems less powerful compared with SIFT and HoG features [36] when used together with SPM [36, 14]. Recent research [21, 3] even shows that the patch-level sparse representation may not be necessary for classification.

1.1. Analysis of Linear Features

The inherent difference of features was extensively studied in this community. But there are still many questions that do not find clear answers. One issue that particularly puzzles many researchers is as follows.

> *Is it possible to make sparse features learned from images more applicable to scene classification?*

We answer this question from the correlation perspective in this paper. We first analyze Independent Component Analysis (ICA) and Sparse Coding (SC) [11], two representatives of linear-model-based feature learning methods, and then present our new framework to improve the discrimination ability of features.

Referring to the illustration in Fig. 1, we select three regions containing the castle tower, lakefront, and flower, from which we extract patches, as shown in (a)-(b). Their structures are completely different. Each patch is then decomposed using the ICA and SC dictionaries shown in (c).

To visualize the statistical regularity of linear responses, we select two pairs of basis vectors, and lay all responses on them for the three visual classes. The distributions for ICA, shown in (e), indicate high correlation among responses. In contrast, sparse coding results that are shown in (f) have their correlation generally degrading to variance due to the employment of over-complete atoms. Both variance and correlation are the second-order statistical regularity among responses. The fact that *their structures vary significantly for different regions* provides a notably useful clue for magnifying the power of linear features in classification.

1.2. Our Method

With the important finding that statistical regularity among linear features embodies *rich* local structure infor-

978-1-4673-1228-8/12/$31.00 ©2012 IEEE

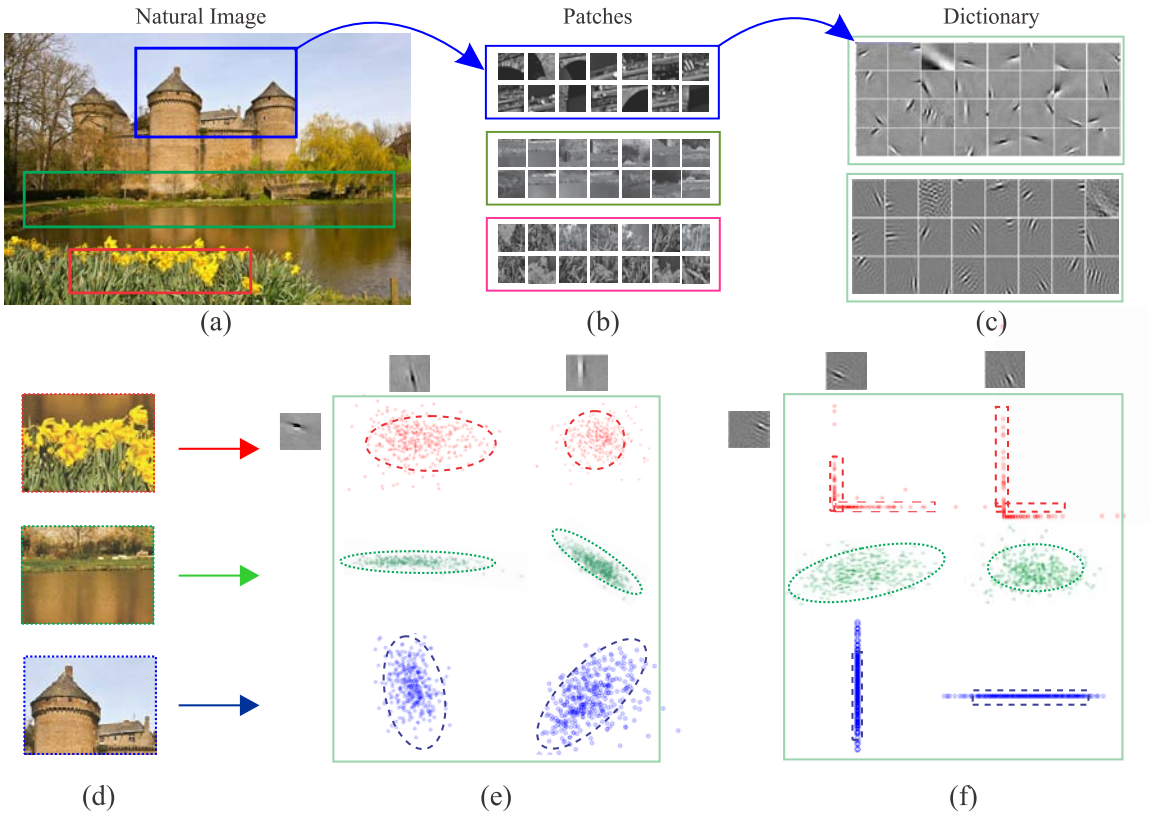

Figure 1. Linear features in scene classification. We select three regions containing the castle tower, lakefront, and flower in (a) and densely extract patches of size 6×6 from them, as shown in (b). Each patch is decomposed using the ICA and sparse coding (SC) dictionaries shown in (c) to acquire linear responses. We plot the distributions of ICA responses on two pairs of basis vectors for the three regions in (e). Similarly, (f) shows the distributions of SC responses on two pairs of basis vectors for the three regions.

mation, instead of purposely reducing or avoiding correlation in linear feature construction, we model correlation explicitly by *combining sparse decomposition of covariance with feature learning*, in order to boost the discrimination ability of linear features. Similar observation exists considering the nonlinearity in complex cells in the primary visual cortex (V1) in computational neuroscience [13].

Our main contribution thus lies on a new model to incorporate covariance patterns into feature construction. We introduce a generative model that captures the covariance patterns from natural images directly. The core idea is to model covariance as sparse linear combination of regular patterns and combine it with learning of linear features in a novel way. With this new representation, inference is accomplished by decomposing the MAP estimation into a few convex optimization problems.

Our method is powerful since it captures the second-order statistics of linear features in natural images. Even working with the most naïve linear features given by the least square (LS) estimator, our method clearly outperforms that with patch-level sparse representation in the SPM framework [21]. Our results are even surprisingly comparable to those generated using the handcrafted SIFT for scene classification [36, 22], as demonstrated in our experiment section. Note that these simple features, when used alone, can only yield rather poor performance.

We use a simple example to demonstrate the effectiveness of our model. We select three categories, i.e., *Forest*, *Coast*, and *Mountain*, from the scene classification dataset *15-scene* [7, 18, 14]. A few examples are shown in Fig. 2. For each image, we collect patches. The linear sparse coding response and our covariance response are shown on bottom left after projecting respective features to two dimensions using the Linear Discriminant Analysis (LDA). It is noticeable that the "LS+COV+SC" features are linearly more discriminative than those with patch-level sparse representation. We note that "LS+COV+SC" is only a toy model. It manifests that our framework can remarkably increase the usefulness of weak linear features in scene classification.

1.3. Related Work

In [12], Karklin and Lewicki proposed learning a linear filter bank that resembles the simple receptive fields by modeling the variance of filter responses. Yu et al. [37] proposed a two-layer sparse coding framework for image clas-

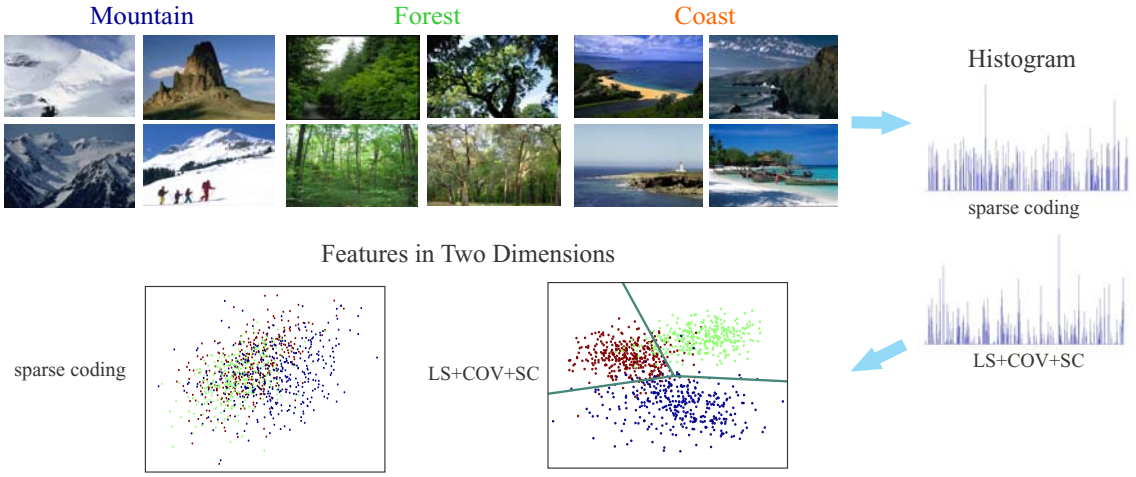

Figure 2. A toy example for scene classification. For each image, we follow the procedure described in Section 4 to collect local regions and patches. A set of basis vectors is learned by running K-means clustering over all patches. Then, we linearly decompose the pathes over the basis by least square (LS) estimation, and calculate the covariance (COV) of the coefficients for each region. By decomposing each covariance into a set of atoms via sparse coding (SC), we get corresponding covariance features. This process is denoted as "LS+COV+SC". For comparison's sake, sparse coding over image patches is also applied. Average pooling is employed on all regions to obtain the feature representation. We project features to two dimensions using Linear Discriminant Analysis (LDA), as shown on bottom left. "LS+COV+SC" is linearly more discriminative than sparse coding over the patches.

sification. These models capture spatial correlation via the variance of sparse coding responses. Learning linear features is still based on the sparsity or independence assumption with only the variance information. In comparison, we directly model covariance of linear features, which encodes pair-wise correlation among feature responses and can work with simpler linear features in scene classification.

Region covariance [27] considers the covariance descriptor of handcrafted features. Sivalingam et al. [24, 25] proposed tensor sparse coding by further decomposing region covariance into a set of linear atoms. Although the term "covariance" is similarly adopted, our model is different by nature given the new generative model that accounts for natural image statistics and simultaneously learns linear features and their covariance patterns.

Hierarchical models capture the higher-order correlation of linear feature by stacking several layers together, where each layer outputs a (non-linear) mapping of its input. The Deep Belief Network (DBN) [8, 1] is a representative in this batch. Convolutional DBN [15] and hierarchial deconvolutional network [38] are the variations.With hierarchies, DBNs work generally better than the one-layer sparse coding along with the cost of the increased number of parameters, higher inference complexity, and longer running time.

2. Model

We begin our description from the generalized form of linear representation. Given a vectorized image patch $x_i \in R^n$, x_i can be represented as the linear combination of a set of pre-learned atoms in dictionary $D = [d_1 \cdots d_j \cdots d_m] \in R^{n \times m}$ along with white Gaussian noise ε:

$$x_i = \sum_{j=1}^{m} \alpha_i^j d_j + \varepsilon = D\alpha_i + \varepsilon, \tag{1}$$

where m is the number of atoms in the dictionary and $A = [\alpha_1, ..., \alpha_i, ..., \alpha_n] \in R^{m \times n}$ is the corresponding coefficient set. Its j-th element is denoted as α_i^j. Given an input patch x_i, the corresponding α_i is also referred to as the linear feature response. With Eq. (1), sparse coding [10] imposes additional independence and sparsity assumption over coefficients α, expressed as

$$p(\alpha) = \prod_i p(\alpha_i), \tag{2}$$

where $p(\alpha_i) \propto \exp(-\lambda|\alpha_i|)$.

Instead of relying on sparse coding to learn α, we propose learning statistical dependency among linear features for classification. An intuitive way to capture correlation among α is to model α as a zero mean multivariate Gaussian controlled by its covariance Σ. Formally, the coefficient α, when conditioned on the covariance Σ, follows a multivariate zero mean Gaussian [6] and is expressed as

$$p(\alpha|\Sigma) = \frac{1}{(2\pi)^{m/2}} \frac{1}{\sqrt{|\Sigma|}} exp\{-\frac{1}{2}(\alpha)^T \Sigma^{-1}(\alpha)\}. \tag{3}$$

An inherent difficulty in (3) is that the covariance matrix lies on the connected Riemann Manifold [28, 27, 24, 9] rather than the Euclidean space. The distance between two covariance matrices has to be defined geodesically, which makes

traditional sparse coding not applicable to covariance matrix decomposition. To address it, we employ the Bregman divergence [32, 24], which models Σ as an inverse-Wishart distribution. The original model thus becomes

$$p(\Sigma|\Theta, l) = \frac{|\Theta|^{l/2}|\Sigma|^{-(m+l+1)/2} exp\{-\frac{1}{2}tr(\Theta\Sigma^{-1})\}}{2^{ml/2}\Gamma_m(l/2)}, \quad (4)$$

where l is the degree of freedom, Γ is the multivariate Gamma function and $\Theta \in R^{m\times m}$ is the reference covariance that can be linearly decomposed into a few simple patterns C_k given by

$$\Theta = \sum_{k=1}^{n_c} \beta_k C_k, \quad (5)$$

where n_c is the number of covariance patterns and $C_k \in R^{m\times m}$ is the positive definite covariance pattern. The coefficient β_k satisfies $\beta_k \geq 0$ and follows a non-negative Laplacian distribution:

$$p(\beta) = \frac{\lambda}{2}\exp(-\lambda\beta). \quad (6)$$

The decomposition procedure is simple and fully captures the correlation of filter responses. It encodes the covariance with basis C_k learned from images, making β a discriminative feature representation for local regions.

3. Learning and Inference

For inference of features, using MAP separately for A and β could induce non-convex optimization. We instead apply joint MAP estimation of A, β, and Σ given that coordinate-wise convex update is possible.

Given a set of patches $X = \{x_i\}$ within a region, dictionary D, and the covariance patterns C, we infer feature β in the MAP framework:

$$\max p(A, \Sigma, \beta|X, D, \{C_k\}) \sim \prod_i (p(x_i|\alpha_i, D)p(\alpha_i|\Sigma))p(\Sigma|\beta, \{C_k\})p(\beta). \quad (7)$$

Taking the negative logarithm and based on the fact [32, 24] that there is an equivalence between modeling the covariance matrix as an inverse wishart distribution and the LogDet divergence, the objective function can be approximated as:

$$\begin{aligned}\tilde{L}(A, \Sigma, \beta) = &\frac{\eta}{2}\|X - DA\|_F^2 + \frac{1}{2}tr(A^T\Sigma^{-1}A) \\ &+ \frac{\mu}{2}\log(\det(\Sigma)) - \frac{\nu}{2}\log(\det(\sum_{k=1}^{n_c}\beta_k C_k)) \\ &+ \frac{1}{2}tr(\sum_{k=1}^{n_c}\beta_k C_k\Sigma^{-1}) + \lambda\|\beta\|_1,\end{aligned} \quad (8)$$

where η is inversely proportional to the variance of Gaussian noise ε in Eq. (1). μ and ν are the corresponding coefficients. We perform the inference via Block Coordinate Descent (BCD) [26]. The basic idea is to iteratively optimize a group of variables while fixing the others.

Solve for A Computing A needs to solve

$$\min_A \quad \frac{\eta}{2}\|X - DA\|_F^2 + \frac{1}{2}tr(A^T\Sigma^{-1}A). \quad (9)$$

By setting the first-order derivative to zero and noting that Σ is symmetric, the solution is in a closed form:

$$A^* = \eta(\eta D^T D + \Sigma^{-1})^{-1}D^T X. \quad (10)$$

Solve for Σ With other variables fixed, the optimization simplifies to

$$\min_\Sigma \frac{1}{2}tr(\Sigma^{-1}(AA^T + \sum_{k=1}^{n_c}\beta_k C_k)) + \frac{\mu}{2}\log(\det(\Sigma)). \quad (11)$$

The solution is obtained by setting the first-order derivative to zero, which yields

$$\Sigma^* = \frac{1}{\mu}(AA^T + \sum_{k=1}^{n_c}\beta_k C_k). \quad (12)$$

Solve for β Given the function simplified to

$$\min_\beta \frac{1}{2}tr(\sum_{k=1}^{n_c}\beta_k C_k\Sigma^{-1}) - \frac{\nu}{2}\log(\det(\sum_{k=1}^{n_c}\beta_k C_k)) + \lambda\|\beta\|_1, \quad (13)$$

we note that the optimization is convex and can be converted to determinant maximization [24], or MAXDET, where interior point solvers exist [29].

For further acceleration, we select a few (5-10) nearest neighbors of Σ in $\{C_k\}$ according to the Euclidean distances, and then solve the simple MAXDET optimization problem without the sparsity constraints on the covariance patterns. This procedure yields

$$\begin{aligned}\min_\beta \quad &\sum_{C_k\in N(\Sigma)}\beta_k tr(C_k\Sigma^{-1}) + \nu\log(\det(\sum_{C_k\in N(\Sigma)}\beta_k C_k)^{-1}) \\ s.t. \quad &\beta \geq 0, \quad \Sigma - \sum_{C_k}\beta_k C_k \succ 0\end{aligned}$$

We use gradient descent to find the result. Similar to [31], the approximation combines locality and sparsity in a highly efficient manner.

3.1. Learning

With feature inference, given the set of regions $\{X^r\}$ that are independently drawn from the images, learning dictionaries D and C can be achieved by solving

$$\max p(D, C_k|\{X^{r_i}\}) = \prod_{r_i} p(D, C_k|X^{r_i}) \sim \prod_{r_i}(p(X^{r_i}|A^{r_i}, D)p(A^{r_i}|\Sigma^{r_i}))p(\Sigma^{r_i}|\beta^{r_i}, \{C_k\})p(\beta^{r_i}). \quad (14)$$

The negative log likelihood function is

$$L = \sum_{r_i} L(A^{r_i}, \Sigma^{r_i}, \beta^{r_i}, D, \{C_k\}), \tag{15}$$

where $L(A^{r_i}, \Sigma^{r_i}, \beta^{r_i}, D, \{C_k\})$ has the same form as the $L(\cdot)$ in Eq. (8).

Online Dictionary Learning The optimal D is given by

$$\min_D f(D) = \frac{\eta}{2} \sum_{r_i} \|X^{r_i} - DA^{r_i}\|_F^2. \tag{16}$$

The function can be solved by a simple least square method. In practice, if there are millions of local regions X^{r_i}, it may not be feasible to put all patches to the memory. We employ an online dictionary learning process, which only updates a small batch B for local regions. The process is expressed as

$$\begin{aligned} D &\leftarrow D - \alpha \nabla f(D), \\ \nabla f(D) &= D(\sum_{r_i \in B} A^{r_i} A^{r_i T}) - \sum_{r_i \in B} X^{r_i} A^{r_i T}, \end{aligned} \tag{17}$$

where α is selected by line search. Since the linear combination is only up to a scale, we normalize all atoms in the dictionary by setting $\|d\|_2^2 = 1$.

Online Covariance Learning We update each pattern sequentially. For each C_k, the optimal solution is

$$\begin{aligned} \min_{C_k} f(C_k) = &\sum_{r_i} (\frac{1}{2} tr(\sum_{k=1}^{n_c} \beta_k^{r_i} C_k (\Sigma^{r_i})^{-1}) \\ &- \frac{\nu}{2} \log(\det(\sum_{k=1}^{n_c} \beta_k^{r_i} C_k))), \\ s.t. \quad &C_k \succ 0 \end{aligned} \tag{18}$$

We also adopt an online process to solve it. Each time, we only read a small batch B of covariance and update patterns incrementally. The process is expressed as

$$\begin{aligned} C_k &\leftarrow C_k - \alpha \nabla f(C_k), \\ \nabla f(C_k) &= \sum_{r_i \in B, \beta_k^{r_i} > 0} \beta_k^{r_i} ((\Sigma^{r_i})^{-1} - \nu (\sum_k \beta_k^{r_i} C_k)^{-1}). \end{aligned} \tag{19}$$

Similarly, α is obtained by line search and we normalize the covariance patterns by setting $tr(C_k) = 1$.

4. Experiments

We evaluate our new model, which couples linear features and their covariance patterns in scene classification. We follow the standard procedure to compute the feature representation.

Patch-level representation Each patch contains 5×5 pixels sampled from a grid with step size 2 (pixels). Patch level representation is obtained by computing coefficients corresponding to our learned 16 linear filters, as shown in Fig. 3(a). They have edge-let shapes.

Region-level representation Each region contains 7×7 patches sampled from a grid with step size 2 (patches). Patch level representation in each region is used to infer the region-level covariance features β (Eq. (13)) based on the 4096 atoms illustrated in Fig. 3(b). We set $\mu = \nu = 1$ for inference.

Image-level representation Given region representation β, the spatial pyramid matching (SPM) with scales 1, 2, and 4 in three levels and max pooling are used to get the feature for the whole image β_{spm}. It is followed by linear SVM for classification.

To get β_{spm}, In each level i of the image pyramid, we divide the region-level features β into i^2 bins. In each bin n, max pooling is performed to acquire $max(\beta(n))$, where $\beta(n)$ is the set of local-region features within the bin n. Finally, the image level feature β_{spm} is simply the concatenation of all bins at all scales.

A subset of 40 covariance atoms from our learned dictionary C (total size 4096) is shown in Fig. 3(b). These atoms exhibit a common pattern – that is, the diagonal elements are bright – indicating that the variance is generally large. The off-diagonal values reveal different levels of correlation among linear filters.

4.1. Structure Mapping

Sparse covariance patterns encode local structure information. Our model, therefore, can be considered as an efficient mapping from the image domain to the structure space. A conjecture is that similar scenes should be "neighbors" to each other in the structure space. To verify it, we conduct a simple experiment on scene retrieval based on the *15-scene* dataset.

We define the Euclidean distance of the sparse covariance patterns for two scene images S_1 and S_2 as the distance in the structure space, that is

$$d(S_1, S_2) \triangleq ||\beta_{spm1} - \beta_{spm2}||_2^2, \tag{20}$$

where β_{spm1} and β_{spm2} are the image features based on our local covariance patterns for S_1 and S_2 respectively. Their construction is described above.

For each randomly picked query image, we compute distance d between it and all other scene images and show the four images with shortest distances in Fig. 4. To visualize local structures, we construct the latent covariance matrices Θ via Eq. (5). Because they are symmetric, we take the upper triangulars and use PCA to obtain the three major principal components for each matrix. These components are mapped directly to three channels in a color space using the method presented in [16]. Specifically, the first component is mapped to channel $R + G + B$; the second is mapped to $R - G$; and the third is mapped to channel $R/2 + G/2 - B$. It

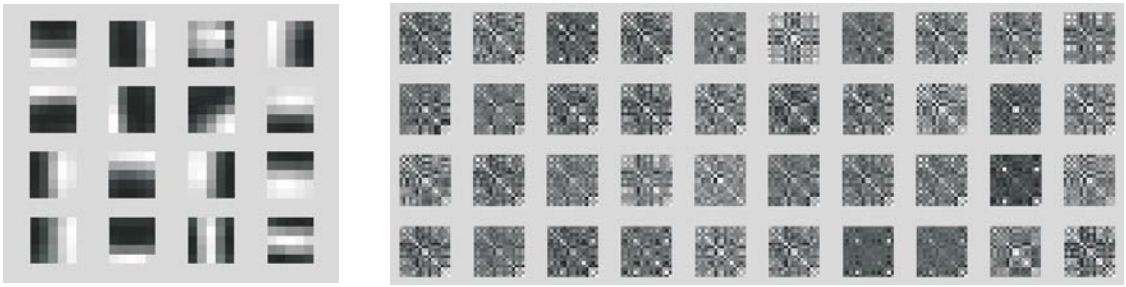

Figure 3. Learned filters by our method on dataset *15-scene*. (a) Atoms in dictionary D. (b) Covariance atoms in dictionary C.

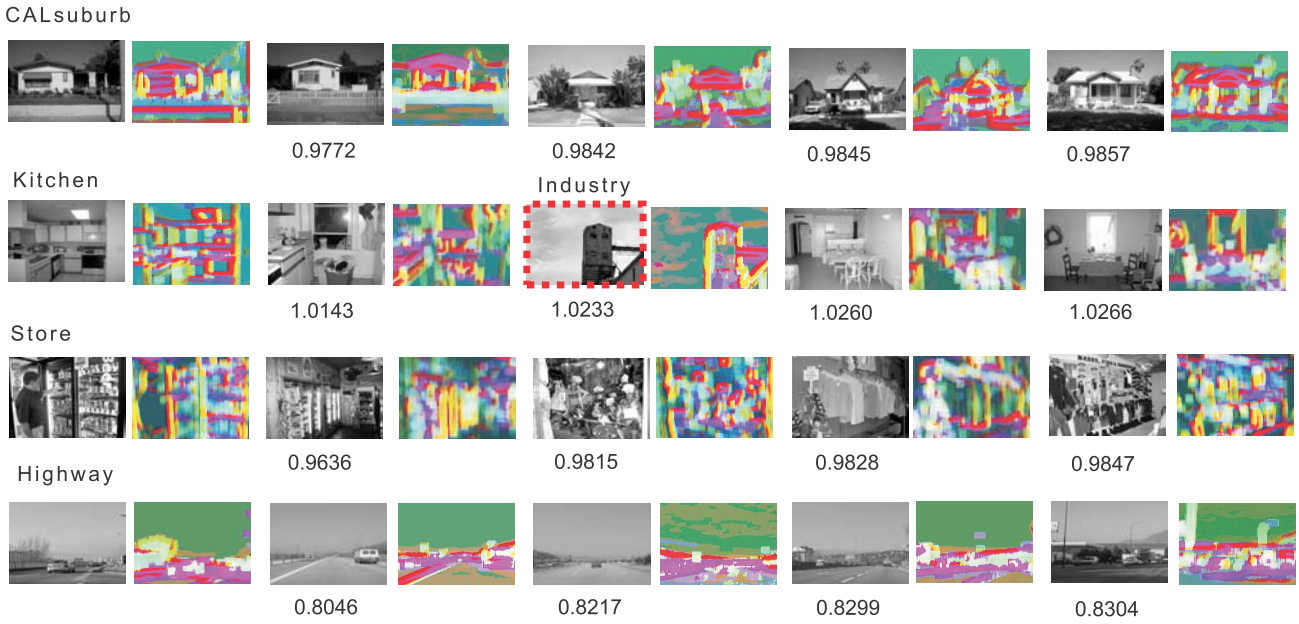

Figure 4. Scene retrieval on the *15-scene* dataset. For each query scene image S, we compute spatial pyramid representation β of local sparse covariance patterns, and take the 4 Nearest Neighbors (NN) measured by Eq. (20) from other 4484 images in the dataset. The query image is the left most one in each row. For each output image, its distance d to the query one is shown. The mismatched one is marked with the red dotted rectangle.

is observable in Fig. 4 that similar local structures have similar color, indicating that our covariance patterns are able to distinguish among local details.

4.2. 15-Scene Classification

We apply our method to scene classification on the *15-scene* dataset [7, 18, 14]. With $200 - 400$ images in each category, the dataset contains 4485 images in total. The average image resolution is 300×250 (pixels). We use 100 images per category for training. All the rest are used for testing. To accurately evaluate the effect of our covariance representation, and to make our method scalable to large-scale data, we only use the simple linear SVM [36] classifier.

In experiments, we randomly collect 4 million patches with size 5×5 from the whole dataset, and learn the dictionary D and C online with the size of 16 and 4096 respectively. Following the feature representation described in Section 4, we decompose each patch linearly over the pre-learned 16 basis vectors. Covariance for each region is then computed based on these linear coefficients, followed by either sparse coding [36] or joint optimization through Eq. (9-13) over the pre-learned over-complete *covariance patterns* C. We denote the process involving sparse coding as "LS+COV+SC" and our joint optimization as "Sparse Covariance Patterns" (SCP).

We take 10 rounds to get the average classification results. In each round, we randomly select the set of training data and leave the rest for testing. Our results are listed in Table 1 with the comparison with the state-of-the-arts [34, 36]. It is noticeable that even the simplest "LS+COV+SC" works reasonably well, compared with powerful handcrafted features. Moreover, our SCP model performs slightly better than sparse coding over SIFT (Sc-

HoG $2*2$ [34]	81.0
GIST [34]	74.7
SSIM [34]	77.2
ScSPM [36] (SIFT+SC+SPM)	80.28 ± 0.93
KSPM [22]	81.40 ± 0.50
"LS+COV+SC"+SPM	79.20 ± 0.32
"SCP"+SPM	80.43 ± 0.49

Table 1. Average classification rates (%) of different methods. Each algorithm is tested for 10 rounds. Our method with the simple linear classifier is comparable with those using handcrafted features and sophisticated nonlinear classifiers.

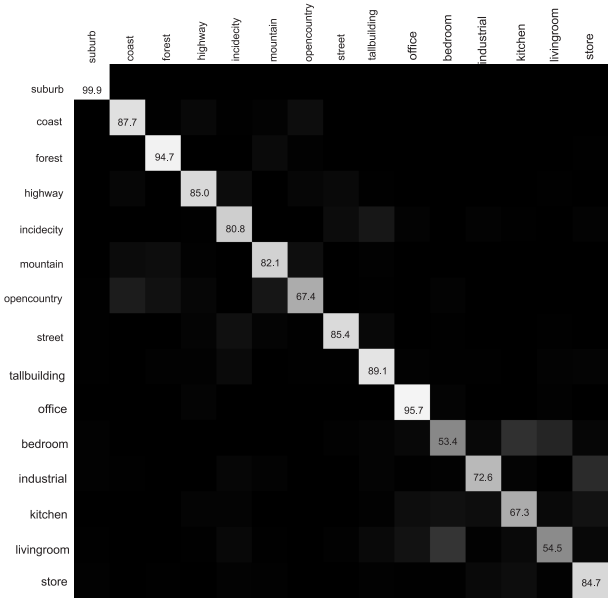

Figure 5. Confusion table for *15-Scene* Classification. The diagonal values are the average classification rates for individual classes while the percentage of images from class i that were misidentified as class j is captured in the i-th row and j-th column of Confusion Matrix.

SPM) [36] and is surprisingly comparable with the kernel SPM [34], which uses non-linear classifiers. We also note that $HoG2*2$, $GIST$, and $SSIM$ are all combined with kernel SVM while our two methods and ScSPM only adopt linear SVM.

We also present the confusion matrix in Fig. 5. The proposed method performs quite well on a few scene categories, including suburb, forest, store, coast, and office. Accuracy falls for bedroom and living room classes. We explain that our method captures local structure information similar to region covariance. Its discrimination power decreases if scene structure variation is very large within some classes.

HoG	22.8
GIST+grayscale	22.0
GIST+Color	29.7
KSPM	34.4
"LS+COV+SC"+SPM	30.1
"SCP"+SPM	33.7

Table 2. Average classification rates (%). The values for HoG, GIST+grayscale, GIST+color, and SIFT+SPM are obtained from [19]. All methods are tested with the same setting as described in [20].

4.3. Indoor Scene Recognition

We also apply our method to the more challenging MIT indoor scene dataset [20]. The difficulty lies in the large variation in both global structures and local details. This dataset contains 67 classes with more than $10K$ images. We follow the training/testing split listed in [20] – that is, for each class, around 80 images are used for training and 20 images are for testing. Again, we only consider the linear classifier, and compare our features with others, including HoG, SIFT, GIST, and color statistics, from the baseline results in [19].

We use the same parameter setting as that for 15-*scene* classification. The dictionaries of D and C_k are also those trained in Section 4.2, which indicates the insensitivity of our method to dictionaries trained in different scene datasets. Our results are reported in Table 2 and are compared with several other methods. SVM with a Gaussian kernel is used in the baseline of "HoG", "GIST+grayscale", and "GIST+grayscale". Our simple "LS+COV+SC" works fairly well and our "SCP" further improves the results with the performance comparable to "KSPM".

5. Conclusion

We have presented a new feature learning framework for scene classification. Based on the observation that statistical regularity of linear representation embodies local structure information, we proposed learning sparse covariance patterns among linear basis vectors. In our model, the sparse constraint is enforced on the second-order correlation for more effective discrimination. Our method explores nonlinear operation on linear features, indicating an intriguing way to boost the power of traditional linear filters for high-level vision tasks.

Acknowledgments

The work described in this paper was supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region (Project No. 413110) and in part by ARO MURI award (No. W911NF-11-1-0046) and

National Science Foundation award IIS-1016772.

References

[1] Y. Bengio, P. Lamblin, D. Popovici, and H. Larochelle. Greedy layer-wise training of deep networks. In *Annual Conference on Neural Information Processing Systems (NIPS)*, 2007.

[2] Y. Boureau, F. Bach, Y. LeCun, and J. Ponce. Learning mid-level features for recognition. In *CVPR*, 2010.

[3] A. Coates and A. Ng. The importance of encoding versus training with sparse coding and vector quantization. In *International Conference on Machine Learning (ICML)*, 2011.

[4] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005.

[5] M. Elad and M. Aharon. Image denoising via sparse and redundant representations over learned dictionaries. *IEEE Transactions on Image Processing (TIP)*, 15(12):3736–3745, 2006.

[6] T. Eltoft, T. Kim, and T. Lee. On the multivariate laplace distribution. *IEEE Signal Processing Letters*, 13(5):300–303, 2006.

[7] L. Fei-Fei and P. Perona. A bayesian hierarchical model for learning natural scene categories. In *CVPR*, 2005.

[8] G. Hinton, S. Osindero, and Y. Teh. A fast learning algorithm for deep belief nets. *Neural computation*, 18(7):1527–1554, 2006.

[9] Y. Hong, Q. Li, J. Jiang, and Z. Tu. Learning a mixture of sparse distance metrics for classification and dimensionality reduction. In *ICCV*, 2011.

[10] K. Huang and S. Aviyente. Sparse representation for signal classification. *Annual Conference on Neural Information Processing Systems (NIPS)*, 2007.

[11] A. Hyvärinen, J. Hurri, and P. Hoyer. *Natural Image Statistics: A probabilistic approach to early computational vision*, volume 39. Springer-Verlag New York Inc, 2009.

[12] Y. Karklin and M. Lewicki. Is early vision optimized for extracting higher-order dependencies? *Annual Conference on Neural Information Processing Systems (NIPS)*, 2006.

[13] Y. Karklin and M. Lewicki. Emergence of complex cell properties by learning to generalize in natural scenes. *Nature*, 457:83–86, 2008.

[14] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR*, 2006.

[15] H. Lee, R. Grosse, R. Ranganath, and A. Ng. Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations. In *International Conference on Machine Learning (ICML)*, 2009.

[16] C. Liu, J. Yuen, and A. Torralba. Sift flow: Dense correspondence across scenes and its applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 33(5):978–994, 2011.

[17] D. Lowe. Object recognition from local scale-invariant features. In *ICCV*, 1999.

[18] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, 42(3):145–175, 2001.

[19] M. Pandey and S. Lazebnik. Scene recognition and weakly supervised object localization with deformable part-based models. In *ICCV*, 2011.

[20] A. Quattoni and A. Torralba. Recognizing indoor scenes. In *CVPR*, 2009.

[21] R. Rigamonti, M. Brown, V. Lepetit, and E. CVLab. Are sparse representations really relevant for image classification? In *CVPR*, 2011.

[22] C. Schmid. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR*, 2006.

[23] T. Serre, L. Wolf, and T. Poggio. Object recognition with features inspired by visual cortex. In *CVPR*, 2005.

[24] R. Sivalingam, D. Boley, V. Morellas, and N. Papanikolopoulos. Tensor sparse coding for region covariances. In *ECCV*, 2010.

[25] R. Sivalingam, D. Boley, V. Morellas, and N. Papanikolopoulos. Positive definite dictionary learning for region covariances. In *ICCV*, 2011.

[26] P. Tseng. Convergence of a block coordinate descent method for nondifferentiable minimization. *Journal of optimization theory and applications*, 109(3):475–494, 2001.

[27] O. Tuzel, F. Porikli, and P. Meer. Region covariance: A fast descriptor for detection and classification. In *ECCV*, 2006.

[28] O. Tuzel, F. Porikli, and P. Meer. Pedestrian detection via classification on riemannian manifolds. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 30(10):1713–1727, 2008.

[29] L. Vandenberghe, S. Boyd, and S. po Wu. Determinant maximization with linear matrix inequality constraints. *SIAM Journal on Matrix Analysis and Applications*, 19:499–533, 1998.

[30] W. Vinje and J. Gallant. Sparse coding and decorrelation in primary visual cortex during natural vision. *Science*, 287(5456):1273, 2000.

[31] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong. Locality-constrained linear coding for image classification. In *CVPR*, 2010.

[32] S. Wang and R. Jin. An information geometry approach for distance metric learning. In *International Conference on Artificial Intelligence and Statistics(AISTATS)*, 2009.

[33] J. Wu and J. M. Rehg. Centrist: A visual descriptor for scene categorization. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 33(8):1489–1501, 2011.

[34] J. Xiao, J. Hays, K. Ehinger, A. Oliva, and A. Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *CVPR*, 2010.

[35] J. Yang, J. Wright, T. Huang, and Y. Ma. Image super-resolution as sparse representation of raw image patches. In *CVPR*, 2008.

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

[37] K. Yu, Y. Lin, and J. Lafferty. Learning image representations from the pixel level via hierarchical sparse coding. In *CVPR*, 2011.

[38] M. Zeiler, D. Krishnan, G. Taylor, and R. Fergus. Deconvolutional networks. In *CVPR*, 2010.