Multi-Level Structured Image Coding on High-Dimensional Image Representation

Li-Jia Li*[1,2], Jun Zhu*[3,4], Hao Su[1], Eric P. Xing[3], Li Fei-Fei[1]

1 Computer Science Department, Stanford University, 2 Yahoo! Research
3 Machine Learning Department, Carnegie Mellon University, 4 Tsinghua University

Abstract. Robust image representations such as classemes [1], Object Bank (OB) [2], spatial pyramid representation(SPM) [3] have been proposed, showing superior performance in various high level visual recognition tasks. Our work is motivated by the need of exploring rich structural information encoded by these image representations. In this paper, we propose a novel Multi-Level Structured Image Coding approach to uncover the structure embedded in representations with rich regular structural information by learning a structured dictionary from it. Specifically, we choose Object Bank [2] to demonstrate our algorithm since it encodes both semantics and spatial location as structural information. By using the learned structured dictionary from Object Bank, we can compute a lower-dimensional and more compact encoding of the image features while preserving and accentuating the rich semantic and spatial information of OB. Our framework is an unsupervised method based on minimizing the reconstruction error of the image and object codes, with an innovative multi-level structural regularization scheme. The object dictionary and the image code obtained by our model offer intriguing intuition of real-world image structures while preserving informative structure of the original OB. We show that our more compact representation outperforms several state-of-the-art representations (including the original OB) on a wide range of high-level visual tasks such as scene classification, image retrieval and annotation.

1 Introduction

Of all the modules for a robust visual recognition system, the design of robust image representation is of fundamental importance and has been attracting many vision researchers. Recently there emerges development of effective image representations such as classemes [1], Object Bank [4], spatial pyramid representation (SPM) [3] and related sparse image representations [5, 6]. While these new image representations demonstrate promising visual classification results, there is still ample space to further explore the rich structural information and potential for high level recognition tasks. Such rich structural information could be representation related to concepts in classmes, objects in Object Bank or spatial locations in SPM. Towards this goal, [4] proposed a supervised method to learn a sparse

*indicates equal contributions.

representation for classification task, which selects the most discriminative objects and responses at specific spatial locations in a class-dependent manner. However, not only the sparsification deteriorates classification performance in some cases, the supervision requires extensive labeling effort and prevents its application for recognition tasks such as annotation. Unsupervised approach, on the other hand, has the advantage of being applicable to unlabeled data and enables the learned representation to be applied to general recognition tasks.

In this paper, we propose a novel unsupervised approach, called Multi-Level Structured Image Coding (MUSIC), to compress a high-dimensional structured data in a much more compact format while preserving the original information. Specifically, we apply our algorithm on OB to demonstrate that our method is capable of representing the high-dimensional OB with a much more compact and semantically interpretable representation called *image code*. Given the OB representation of a set of images, MUSIC learns a set of bases that span a lower-dimensional semantic space in which unknown images can be encoded with compact image representation, by minimizing a regularized reconstruction error over the input representation. Specifically, we make the following contributions:

1. We demonstrate to compress the original OB feature into a much lower dimensional latent space (40 folds reduction) while preserving the rich semantic and spatial information encoded in the feature.
2. We propose a two-layer structured coding scheme at both object and image levels; and a *structured object dictionary* that consists of both unique bases that encodes the canonical spatial distribution patterns of every specific object in OB, as well as shared bases that are generic to all objects. An efficient coordinate descent algorithm is developed to solve the optimization problem of image coding and dictionary learning in a fully unsupervised fashion.
3. The resulting image code can be efficiently applied to high level visual tasks such as image classification, retrieval and annotation with superior performance.

2 Related Work

Image representation research has achieved substantial progress in image recognition tasks recently with the emergence of robust representations such as [7], classemes [1], and Object Bank [2] by encoding images as structural composition of semantically meaningful intermediate representations. However, little has been done to uncover the structural information within these rich intermediate representations. [4] explore sparsity within the high dimensional OB by using a supervised approach. In Sec.3.6, we demonstrate the fundamental difference between our method and [4]. Compared to [4], our algorithm explores to learn a low dimensional latent space in an unsupervised manner via regularized projection, which generalizes knowledge to unseen classes.

One notable method for obtaining unsupervised features is deep learning algorithms [8–10]. Deep belief network (DBN) can be viewed as composition of multiple levels of non-linear operations. Sparse coding methods (SPC) [11, 12, 6, 13–17] have shown impressive potential in different applications. MUSIC fundamentally differs from these methods. First, MUSIC is a two-layer regularized projection method, which explicitly leverages the rich semantic and spatial information in OB to achieve compactness at both object and image levels. Second,

MUSIC learns a structured dictionary by explicitly defining object-specific and shared bases. The usage of object-specific and shared bases is related to [17], which learns specific/shared dictionary elements for multi-view data analysis. We defer the detailed comparison to Sec.3.6 after we introduce our model.

3 Multi-Level Structured Image Coding

In this section, we describe our multi-level structured image coding algorithm by using Object Bank as an example. We first summarize the Object Bank (OB) representation introduced by Li *et al.* [2]. A set of response maps are first obtained by applying a set of pre-trained object detectors at multiple scales in an image to extract the response value from each pixel at each scale. The resulting response map at each scale is then divided into grids similar to SPM [3]. Within each grid, maximum response score for each object filter is selected to build the final OB representation[1]. Let O denote the number of object filters and G denote the total number of grids for an image ($G = nScales \times nGridsperscale$ i.e. $12\times(1+4+16)=252$). An OB representation is constructed by concatenating a set of O *object-wise* sub-vectors $\mathbf{x}=[\mathbf{x}_1;\cdots;\mathbf{x}_O]$, where $\mathbf{x}_o \in \mathbb{R}^G$ denotes the responses of the oth object filter across all grids[2]. The overall dimensionality of the OB representation is $O\times G$, which can easily grow into tens of thousands as the number of object filters increases.

The OB representation departs from traditional low-level image features by encoding both semantic and spatial information. For example, the response maps of 'sky' or 'airplane' would show stronger signals in the upper half of the image, whereas 'beach' or 'car' would be the opposite. Furthermore, the OB representation is an over-complete characterization of the image, in which redundant information can be across different scales, overlapping spatial regions, and even at object levels. The redundancy in OB representation potentiates a robust compression without scarifying useful information. To extract a low dimensional compact representation from the over-complete OB while preserving the semantic and spatial information, we base our model upon regularized linear projection. We present the *Multi-Level Structured Image Coding* (MUSIC) for unsupervised learning of a structured dictionary constituted by a set of bases spanning the low dimensional latent space. With this dictionary, the high dimensional OB representation can be encoded by using the bases in the low dimensional latent space to generate the compact image code with high fidelity.

In a typical encoding process, an input data $\mathbf{x}$ can be represented as a linear combination of K basis vectors, $\mathbf{x} \approx \sum_{k=1}^{K} s_k \beta_k$ learned via a loss minimization scheme $\min \sum_d \|\mathbf{x}^{(d)} - \beta \mathbf{s}^{(d)}\|_2^2$, where d represents the data index. The weight vector $\mathbf{s} \in \mathbb{R}^K$ is called a *code*. The set of bases $\beta = [\beta_1;\cdots;\beta_K]$ is the *dictionary*.

Recent sparse coding (SPC) approaches [18, 11, 12] apply regularizers to enforce sparsity, generating the so called *sparse code* usually with very few non-zero

[1] We follow [2] and use the same 177 detectors, 12 scales and 3 pyramid levels.

[2] There can be an alternative representation, in which $\mathbf{x}$ is organized as a concatenation of set of *grid-wise* sub-vectors, where each sub-vector $\mathbf{x}_g \in \mathbb{R}^O$ denotes the responses produced by all the object filters in grid g. For simplicity, we focus on the object-wise concatenation as an example.

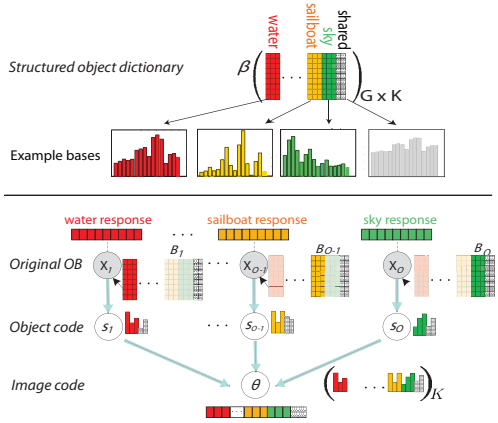

Fig. 1. Top: The learned structured object dictionary β. In β, the colored grids (column-wise) represent object-specific bases while the shaded grids are shared bases. Across all variables, grids in the same color are directly correlated. We show one example basis for each object and one shared basis. Bottom: An illustration of MUSIC for inferring a compact image code from high-dimensional OB features. Here, $\mathbf{x}_o \in \mathbb{R}^G$ is the response of an object filter in the original OB, $\mathbf{s}_o$ is the code of object o whose dimension is much lower than G (Sec. 3.1), θ is the image code (Sec. 3.2) that aggregates $\mathbf{s}_o$ to achieve a single compressed representation for entire image. B_o represents all the bases needed for reconstructing signals from object o, in which faded columns represents unused bases. (Best viewed in color and with magnification.)

elements. SPC can be directly applied to the OB features, treating the entire $O \times G$ dimensional OB representation $\mathbf{x}$ of an image as an input sample, from which the latent-space representation can be computed using some matrix factorization methods. Let's refer this method as basic SPC. The problem is that the basic SPC approach can be extremely inefficient because all the bases need to be as high-dimensional as the original OB. Moreover, this method does not consider the rich structural information in OB discussed earlier.

Here, we seek to obtain a low dimensional projection θ of the OB features $\mathbf{x}$ in a latent space spanned by a *structured dictionary* learned efficiently with MUSIC by exploring the structural information in OB. Fig. 1 shows MUSIC, which contains an object-level for object-wise encoding and an image-level for image-wise encoding. Below, we elaborate our multi-level coding scheme.

3.1 Object Coding in MUSIC

To avoid expensive high-dimensional matrix factorization, MUSIC performs encoding at *object-level* (object coding) for individual object-wise response subvector, i.e., $\mathbf{x}_o$ in $\mathbf{x}$, instead of the whole input $\mathbf{x}$. Given an input $\mathbf{x}$ resultant from a concatenation of O subvectors, each corresponding to an object-specific spatial response, we begin by reconstructing each subvector $\mathbf{x}_o$ from some bases in a dictionary, as shown in the first layer from $\mathbf{x}$ to $\mathbf{s}$ in Fig. 1. Let β denote a *structured dictionary*, consisting of O sets of object-wise *unique* bases, each set denoted by $\beta_o \equiv \{\boldsymbol{\beta}_{o,1}, ..., \boldsymbol{\beta}_{o,M}\}$ where M is the number of bases unique to every object; and one set of L *shared* bases $\beta_c \equiv \{\boldsymbol{\beta}_{c,1}, ..., \boldsymbol{\beta}_{c,L}\}$ that are used to

reconstruct all the object-wise inputs. Putting all the bases together, we obtain the structured dictionary β as a $G \times K$ matrix, where $K = (O \times M + L)$. Each basis $\boldsymbol{\beta}_k \in \mathbb{R}^G$ in the dictionary roughly represents a canonical response pattern of one object detector on different spatial locations and scales in an image. Ideally, the object-wise unique bases capture private object appearance and spatial patterns of an object whereas the shared bases contain common patterns of all objects. We adopt a linear scheme for object-signal reconstruction:

$$\mathbf{x}_o \approx \sum_{j=1}^{M} u_{o,j} \boldsymbol{\beta}_{o,j} + \sum_{j'=1}^{L} v_{o,j'} \boldsymbol{\beta}_{o,j'} = \beta_o \boldsymbol{u}_o + \beta_c \boldsymbol{v}_o, \tag{1}$$

where $\boldsymbol{u}_o$ and $\boldsymbol{v}_o$ represent the vectors of weights measuring the contributions of the unique bases and shared bases, respectively. We denote $\mathbf{s}_o \equiv \text{cat}(\boldsymbol{u}_o; \boldsymbol{v}_o)$, which is a concatenation of $\boldsymbol{u}_o$ and $\boldsymbol{v}_o$, as the *code* of object o; likewise, $B_o \equiv \text{cat}(\beta_o; \beta_c)$[3] represents all the bases needed for reconstructing the response signals from object o.

Naively, one can code the signals $\mathbf{x}_o^{(d)}$ of image d by estimating $\mathbf{s}_o^{(d)}$ via a loss minimization scheme. To avoid manual engineering for selecting the best number of bases for each object, we enforce a sparsity-inducing regularizer upon this scheme to discover the number of useful bases automatically from a potentially large number of candidates:

$$\{\mathbf{s}_o^{(d)}\} = \arg\min \sum_d \|\mathbf{x}_o^{(d)} - B_o \mathbf{s}_o^{(d)}\|_2^2 + \rho \sum_d \|\mathbf{s}_o^{(d)}\|_1, \tag{2}$$

where ρ is a non-negative constant that balances the regularization term and the reconstruction error term. For different objects, the B_o's contain a common element β_c that couples different codes for similarity within different objects, which may render all $\mathbf{s}_o$'s of a single image not independent of each other due to $\boldsymbol{v}_o$ contained. Therefore, in the sequel we need to furthermore consider a global *image coding* built on the object coding.

3.2 Image Coding in MUSIC

To capture the correlations among the codes of different object-wise responses, MUSIC employs an additional layer of coding as shown in Fig. 1, which aggregates the codes $\mathbf{s}_o$ of all object-wise sub-vectors to achieve a single compressed OB representation θ of the entire image.

We define *image code* θ as a $(O \times M + L)$ dimensional vector of the form: $\theta \equiv \text{cat}(\theta_1; \ldots; \theta_O; \theta_c)$, of which the elements $\theta_o, o = 1, \ldots, O$ correspond to the *prototype code* of *each* of the unique portion $\boldsymbol{u}_o$ of the object code $\mathbf{s}_o$; and the last element θ_c corresponds to the *prototype code* underlying *all* of the shared portion $\boldsymbol{v}_o$. Similar to the object code extraction from $\mathbf{x}_o$'s, we adopt a regularized loss minimization scheme to extract θ from $\mathbf{s}_o$'s:

$$\{\theta\} = \arg\min \sum_{o=1}^{O} \|\theta_o - \boldsymbol{u}_o\|_2^2 + \|\theta_c - \boldsymbol{v}_o\|_2^2 + \lambda \sum_{i=1}^{O+1} \|\theta_i\|_2, \tag{3}$$

where λ is another regularization constant. The choice of loss function determines how image code θ is aggregated from the individual $\mathbf{s}_o$. The square error

[3] Here cat() denotes a column-wise concatenation.

loss above yields an average pooling for shared components and a re-weighted feature concatenation for object-specific components. Since not all object filters are equally important for describing an image, we employ an $\ell_{1,2}$-norm regularizer over the image code for automatically selecting useful object filters[4].

3.3 Structured Dictionary in MUSIC

As mentioned earlier, the *dictionary* β is a $G \times K$ matrix, where $K = O \times M + L$ is the total number of bases, and we use B_o to denote the sub-matrix $\text{cat}(\beta_o; \beta_c)$ for constructing the signals from object o. Fig. 1 illustrates the structure of such a dictionary. We can see that an object-specific basis tends to represent a canonical spatial pattern for a particular object and a shared basis tends to capture the common spatial pattern of all the objects. Therefore, object-specific bases are sharp and vary significantly from one object to another, while a shared basis is much flatter. In MUSIC, the dictionary β is learned in conjunction with the coding process that renders $\mathbf{s}_o$ and θ. This dictionary encodes rich structural information in OB, upon which the compact image code can be extracted. We will present a closer examination of β in Sec. 4.1.

3.4 The MUSIC Model

Putting the three components above together, in order to learn an optimum dictionary β and infer the optimal coding coefficients $(\mathbf{s}, \theta)$, we define MUSIC as a coding/learning scheme based on minimizing a regularized square reconstruction error. Formally, given a set of images $\{\mathbf{x}^{(d)}\}$, we solve the following problem:

$$\min_{\theta,\mathbf{s},\beta} \sum_d L(\mathbf{x}^{(d)}; \mathbf{s}^{(d)}, \beta) + \gamma L(\mathbf{s}^{(d)}, \theta^{(d)}) + \rho\Omega(\mathbf{s}^{(d)}) + \lambda\psi(\theta^{(d)}), \text{s.t.} : \beta \in \mathbb{B}, \tag{4}$$

where $L(\mathbf{x}^{(d)}; \mathbf{s}^{(d)}, \beta) = \sum_o \|\mathbf{x}_o^{(d)} - B_o \mathbf{s}_o^{(d)}\|_2^2$ is the square error between input features and their reconstructions; $L(\mathbf{s}^{(d)}, \theta^{(d)})$ is a similar square error between object codes and image code as defined in Sec. 3.2; $\Omega(\mathbf{s})$ is the ℓ_1-norm of object codes as in problem (2); and $\psi(\theta)$ is the $\ell_{1,2}$-norm of the image code as in problem (3). Here, (λ, γ, ρ) are pre-specified non-negative hyper-parameters, which can be chosen via cross validation. To make the problem identifiable, we put a constraint on the dictionary β. We define $\mathbb{B} = \{\beta : \sum_k \max_j |\beta_{kj}| \leq C\}$, which constrains the $\ell_{1,\infty}$-norm of β to be less or equal to a threshold C [19]. $\ell_{1,\infty}$-norm encourages some of the bases to be entirely zeroed-out. It effectively avoids the spread of shared bases and bias the latent space towards being compact. Intuitively, we expect that each object has fewer canonical spatial patterns.

In summary, MUSIC has the following main innovations over the basic SPC: **1**) Rather than computing a global code directly from the entire input $\mathbf{x}$, it computes sparse codes at *object-level* for each subvector $\mathbf{x}_o$ in $\mathbf{x}$ to avoid expensive high-dimensional matrix factorization. **2**) Rather than using a universal set of

[4] Note that, the $\ell_{1,2}$-norm regularizer over θ can encourage joint sparsity of weights within an object code over all regions, i.e., all elements in a subvector θ_o or θ_c is shrunk to zero simultaneously, which is a desirable bias to clean up spurious object filters. It is also possible to explore other structured sparsity, such as regional effects, but for simplicity, we leave these enhancements to the future work.

bases for reconstructing all $\mathbf{x}_o$'s, it employs a structured dictionary consisting of both (small) basis-sets unique to each object and a basis-set shared by all objects to reconstruct every $\mathbf{x}_o$ for high fidelity and effective extraction of semantic and spatial pattern. **3**) Rather than obtaining a reconstruction simply by pre-specifying the dictionary size, we impose both object-level and image-level sparsity-inducing bias. This enables us to incorporate structural knowledge such as preferred co-occurrence of objects, or appearance and/or filter-response, directly to the image code, which is impossible in the basic SPC described above.

3.5 Optimization Algorithm: Coordinate Descent

Algorithm 1 Dictionary Learning

Input: image corpus $\{\mathbf{x}^{(d)}\}_{d=1}^{D}$, regularization constants (λ, γ, ρ), basis numbers (M, L).
Output: dictionary β
repeat
Coding: *infer the sparse object codes* $\mathbf{s}$ *and image code* θ *for each image using Alg. 2*
Dictionary Learning: *solve the following convex problem for* β *with projected gradient descent*

$$\min_{\beta} \sum_{do} \|\mathbf{x}_o^{(d)} - B_o \mathbf{s}_o^{(d)}\|_2^2, \quad \text{s.t.: } \beta \in \mathbb{B}. \tag{5}$$

until convergence

Algorithm 2 Coding for Compact image code

Input: an image $\mathbf{x}$ and object dictionary β, regularization constants (λ, γ, ρ), basis numbers (M, L).
Output: the image code θ and object codes $\mathbf{s}$.
repeat
for $o = 1$ **to** O **do**
Solve the convex problem (2) for object-code $\mathbf{s}_o$.
end for
Solve the convex problem (3) for θ.
until convergence (Image index is ignored for notation simplicity)

We present an efficient procedure to solve problem (4). We note that the objective function is not joint convex, but it is bi-convex, that is, convex over one of $(\theta, \mathbf{s})$ and β when the other is fixed. One natural choice is the coordinate descent algorithm, which has been widely used in sparse coding [20, 6]. The algorithm alternates between two steps of coding and dictionary learning, as outlined in Alg. 1.

Coding: this step solves problem (4) for $(\theta, \mathbf{s})$ with β fixed, which is a convex problem. We adopt a coordinate descent strategy to iteratively solve it over θ and $\mathbf{s}$, which has been shown suitable for ℓ_1-norm and $\ell_{1,2}$-norm regularized least squares loss [20]. Given the independence assumption of different images, we can

perform this step for each image separately, as outlined in Alg. 2. Here, both problems (2) & (3) have closed-form solutions. The time complexity for solving problem (2) is $\mathcal{O}(O \times (M+L) \times G)$ and problem (3) $\mathcal{O}(O \times (M+L))$. We defer the details to the Appendix.

Update dictionary: this sub-step involves solving the convex problem (5) with a quadratic objective. We solve this problem via spectral projected gradient descent, and the projection to the $\ell_{1,\infty}$-ball is efficient [19][5].

3.6 Comparison with [4] and [17]

Our MUSIC model described above infers a low dimensional latent space of OB based upon regularized linear projection, which requires no supervision. [4] explores the sparsity resides in OB by using regularized logistic regression, where supervision is necessary. Our method generalizes knowledge to unseen classes and copes effectively with scarcity of labeled data, which is a major challenge in real-world applications. In Sec.4.2, we apply our representation to general high level recognition tasks such as retrieval and annotation whereas the representation obtained by [4] is only applicable to classification.

Comparing to [17], which softly couples the dictionaries on multi-views by using a structured regularizer, MUSIC is a multi-level model, which first explicitly defines object-specific and shared bases and then enforces bases selection. Although [17] could potentially be more flexible in identifying the shared and private bases, they are computationally much more demanding to obtain the comparable number of bases. By exploring the structure of OB features and explicitly defining the specific and shared bases, MUSIC can dramatically reduce the computational burden. For example, if we aim to learn M specific plus L shared bases for each object, suppose each object-wise input is one view, [17] needs to learn $O \times (O \times M + L)$ bases in total, while we need only $O \times M + L$ bases. Given that there are hundreds and potentially thousands of objects in the high-level image representations such as OB, MUSIC demonstrates much more potential in scalability.

4 Experiment

In this section, we analyze a number of important properties of the MUSIC approach and evaluate its performance on several high-level recognition tasks, including scene classification, image retrieval and annotation.

4.1 Analysis of MUSIC

Before showing performance of the proposed MUSIC approach on several benchmark applications, we examine some basic properties of MUSIC in this section. We focus on *interpretability* of the learned structured dictionary; *information content* of the image code, and *discriminability* of the MUSIC output over competing methods. We use the UIUC-Sports event dataset, which contains images from 8 event categories. 70 images from each class are used to learn the dictionary and compute the image code. The original OB is generated from 177

[5] More efficient solver for this problem exists, e.g. FISTA [21].

object detectors following [2]. As described in Sec. 3, the overall dimensionality is $O \times G = 44604$. In MUSIC, 7 object specific bases for each object and 1 shared basis are used, which generates the image code with dimensionality of $7 \times 177 + 1 = 1240$, approximately 40-fold reduction from the original OB. The compact image code reconstructs the original OB with high fidelity, which is reflected by average mean square error 0.0052 (less than 0.3%) with standard deviation of 6×10^{-7} in each dimension with multiple randomly initialized runs.

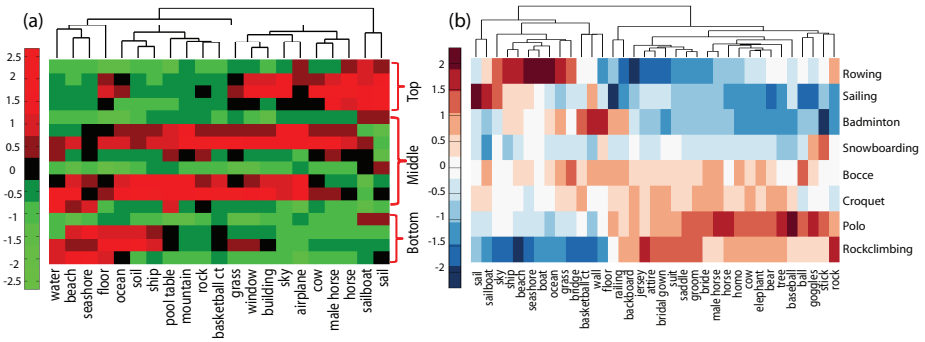

Fig. 2. (a) A heat matrix of the bases in the structured dictionary. Each column corresponds to a basis, and each row corresponds to a spatial location (i.e., a grid), which are grouped as 'Top', 'Middle' and 'Bottom' locations in the image. Object names are displayed at the bottom. A high value of a basis-element in a row indicates that the object is likely to appear in the corresponding grid. The values of each object basis are standardized for salient visualization. Bases are ordered as leaves of a hierarchical clustering of the columns for salient visualization. (b) A heat matrix of the average image code θ of images from different classes are displayed on the right.

A Close Examination of the Basis Dictionary As mentioned earlier, MUSIC learns a structured dictionary that includes O subsets of object-specific bases, and 1 subset of shared bases. Each basis captures the canonical pattern of either an object-specific or a universal OB signal in all the regions (i.e, grids) of the image. Fig. 2(a) shows a few examples of the learned object-specific bases (each as a column in the heat matrix). As we can see, there is an apparent spatial preference revealed in different bases. Bases corresponding to objects such as 'water', 'beach', 'seashore', etc. contain strong (i.e., red) signals at bottom; whereas 'maintain', 'pool table', 'rock', etc. tend to have stronger signals in the middle regions of the image. A simple hierarchical clustering of the bases can reveal semantically coherent grouping based on the spatial preferences of objects; for example, 'water', 'beach' and "seashore" fall into a cluster on the left, whereas 'sail' and 'sailboat' are in the same cluster on the right. However, as spatial preferences are not the only cue for grouping objects (e.g., the apparent heterogeneous composition of the cluster in the middle), we are cautious about over-interpreting this groupings.

A Close Examination of the Image Code We expect the image code inferred by MUSIC from OB to bear sufficient content so that they lead to high-quality reconstruction of the original OB, as well as being semantically interpretable.

Fig. 2(b) shows a heat matrix of subvectors of the average image code obtained by extracting the object prototype code θ_o of 34 example objects for the 8 classes. Here, each row corresponds to an average image code for images from a category, and each column corresponds to an object-specific basis used. Again, columns are ordered by a hierarchical clustering simply for easy visualization. It can be seen that different image categories do exhibit preferred usage of different object-specific bases, reflecting more frequent occurrences of the corresponding objects in the images. For example, 'sail' and 'sailboat' have higher image code values in the 'sailing' class, while 'sky', 'ship' and 'boat' have higher image code values in the 'rowing' class. Such content preference in the image code implies its potential in semantic-based discrimination, as we explore later.

Predictive Performance Now we dissect the building blocks of MUSIC and examine their influences on the predictive power of the inferred image code. The evaluation is based on a scene classification experiment on the UIUC sports data. We use 70 images for training and 60 for test from each class as the default setting for all experiments on this data. In the following, if not specified, we employ a multi-class linear SVM as the default classifier for different image representations, including image code. We compare with the following alternatives:

1. SPC: basic sparse coding that uses all shared bases to compute object codes separately, which were subsequently concatenated into a whole image-level representation[6].
2. specSPC: image code from MUSIC, but using only object-specific bases (i.e., $L = 0$).
3. PCA: a representative dimensionality reduction method.
4. L1-LR [4]: ℓ_1-norm regularized logistic regression (LR) trained directly on the high-dimensional OB representations.
5. OB-SVM [2]: a linear SVM learned from the original OB.

Table 1 summarizes the classification accuracy of different methods. For MUSIC, the dimension of the image code is 1240, much lower than that of the original OB (44604). For SPC, specSPC and PCA, we use 7 bases per object, which gives approximately the same dimensionality. We observe that the full MUSIC outperforms other algorithms. Specifically, the superior performance of MUSIC over specSPC demonstrates that shared bases can help separate common background information from the more semantic salient information regarding unique objects.

Method	**Accuracy**
OB-SVM	77.9%
L1-LR	76.2%
PCA	77.2%
SPC	78.7%
specSPC	78.0%
MUSIC	**81.8%**

Table 1. Classification accuracy of different models.

[6] If we use average or max-pooling in the SPC to obtain image-level representations, it is found that the performance will drop dramatically when only using a modest number (e.g., hundreds) of bases. Using a large number of bases could help but it is much more expensive than MUSIC. For example, suppose we use the same number of bases as that of objects in both SPC and MUSIC (i.e., $M = 1$ & $L = 0$). Then, MUSIC will be roughly the number of objects times faster than SPC. This is because SPC uses all the bases to reconstruct each object-wise OB feature, while MUSIC uses only 1 basis to reconstruct each object-wise OB feature.

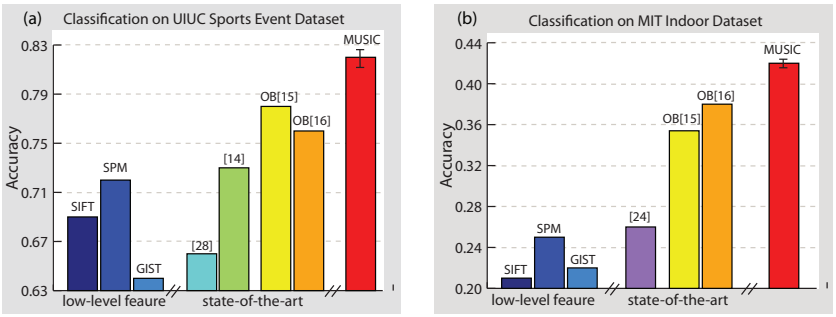

Fig. 3. (a) Comparison of classification performance to the methods that use existing low-level representations and state-of-the-art approaches including previous OB related methods on UIUC sports data. Average accuracy of a multi-way classification is used as the evaluation metric. (b) Comparison of classification performance to the methods that use existing low-level representations and state-of-the-art approaches on MIT Indoor. Error bars represent standard deviation of 5-fold random split of the training/test data.

4.2 Applications in High-level Image Recognition

In this section, we evaluate the potential of image code in high-level visual recognition tasks, specifically: scene classification, image retrieval and annotation.

Scene Classification We first analyze the predictive power of the image code learned by MUSIC for classifying scene images from two complex scene datasets – UIUC sports event [22] and MIT indoor scene [23]. For MIT indoor scene dataset, we follow the settings in [23], using 80 images from each of the 67 classes to train a multi-class linear SVM and test on a total of 1340 images (20 per class). We compare image code obtained by MUSIC with those methods using low-level features (e.g., SIFT, SPM and GIST) and the state-of-the-art algorithms. We use a linear SVM classifier for SIFT and GIST features and a more complex classifier (i.e. SVM with an intersection kernel) for SPM as in [3]. Fig. 3 shows

Method	**MUSIC**	MUSIC-kNN	MUSIC-LR	Self-Taught
Accuracy	**81.8%**	69.5%	79.2%	80.4%

Table 2. Classification performance of different classifiers and self-taught learning (learn on MIT indoor data and apply it to infer image code for images in UIUC sports data) on the image code inferred by MUSIC.

the accuracy of different methods. The improvement of image code from MUSIC over the low-level representations and the state-of-the-art approaches indicates image code can successfully preserve the rich structure and semantic meaning of the OB representation. It is worth noticing that all OB unrelated state-of-the-art algorithms here require extensive supervision during training ([24] and [22] use object labels within each training image. [23] requires manual segmentation of a subset of training images.), whereas MUSIC does not require such supervision. The fact MUSIC outperforms the original OB underscores the importance

of obtaining a more compact feature, where discriminative information in both semantic and spatial domains are preserved, but the smaller dimensionality curtails the high-dimensionality challenge posed by the original OB.

We also investigate how different end classifiers affect the classification performance by using our image code from MUSIC. Here, we compare linear SVM (the default classifier) with kNN and logistic regression (LR), which are denoted by MUSIC-kNN and MUSIC-LR, respectively. We also report the performance of using our image code for self-taught learning [13]. As shown in Table 2, the linear LR and SVM perform comparably. Although inferior to SVM or LR, kNN is comparable to the best method that uses a low-level representation (e.g., SPM [3]) and the state-of-the-art methods as shown in Fig. 3(a).

Our image code is much more efficient when applied to high level visual tasks. Taking scene classification on MITIndoor dataset as an example, training and test efficiency has been improved over 20 and 60 times respectively over the OB representation. Given that coding time for an unknown image is negligible, our representation has much more potential in scalability.

Image Retrieval We investigate the usefulness of our image code inferred from MUSIC on content based image retrieval task, i.e. use a query image to retrieve relevant images[7].

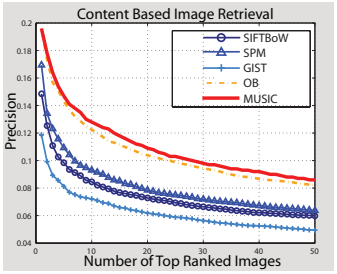

Fig. 4. Content based image retrieval: precision of the the top ranked images by using GIST, BOW, SPM, original OB, and image code on the UIUC sports event dataset. Cosine distance is used as the distance measurement. Best viewed in colors and magnification.

We compare the retrieval performance of image code to that of low-level representations and the original OB [2] on the UIUC sports data, where 130 images from each of the 8 classes are used. We use precision of the top ranked images as the evaluation criterion (Fig. 4). Image code by MUSIC outperforms those using low-level representations with a large margin. We attribute this advantage of MUSIC to its encoding of rich semantic and spatial information. It is worth noticing that although achieving comparable performance, the image code has a much lower dimension (1240) and hence more efficient for practical applications than the OB representation (>40k dimension).

[7] We also test our image code on concept based image retrieval, i.e. use a concept to retrieve images. The compactness of image code makes it a better choice over OB [2] for large scale retrieval application. Image code significantly outperforms the state-of-the-art algorithms in [1], which is largely attributed to the incorporation of the spatial patterns of the objects. See result and details about this experiment in the Appendix.

Image Annotation Our last experiment is to apply the image code to an image annotation task, where a list of image or object concepts is inferred for an image. We conduct this experiment on the UIUC sports data, with 70 images per class for training (18 of them are used as the validation set) and 60 for testing. We train an SVM classifier of each object based on our representation of the images. Given a query image, the classifier makes an annotation prediction for each concept. Our result (48.27% in standard F-measure, used in [24]) is superior than the reported state-of-the-art performance of 38.20% in [24]. We attribute this improvement to the rich semantic information encoded by the object filters, which is suitable for high-level visual tasks such as image annotation. Our method also outperforms the original OB representation (45.46%), suggesting that the compact image code successfully preserve the semantic contents of the image while discarding the noise and redundancy.

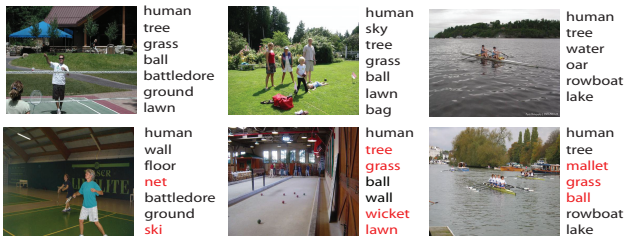

Fig. 5. Example image annotation results by MUSIC. Proposed tags are listed on the right side of the image. Incorrect tags are highlighted in red. The average number of tags proposed is ∼10. For those images with more than 7 tags predicted, only the top 7 tags with highest empirical frequencies in the tag list of that image are shown.

Fig. 5 shows a few example results annotated by MUSIC. One source of mistakes is due to semantic confusion, e.g. 'net' in 2nd row and 1st column is proposed as a result of its expected occurrence in badminton images. Another source of errors is the object filters in OB, such as incorrectly labeling 'ski' as 'stick'. These observations point out several useful future directions in improving our work.

5 Conclusion

We have proposed a novel MUSIC model that learns a structured object dictionary in an unsupervised manner using a high-level representation (i.e., OB) and infers much more compact image code representation (∼ 1k dimension) than the original OB (44604 dimension). Our analysis demonstrates that the structural regularity in the learned dictionary and the inferred image code are consistent with human knowledge. Using the inferred image code, superior performance over original OB can be obtained with a much lower computational cost on various high-level image recognition tasks. We plan to explore the potential of the compact image code in large scale recognition problems, which are infeasible for the original OB due to its high dimensionality and over-completeness.

References

1. Torresani, L., Szummer, M., Fitzgibbon, A.: Efficient Object Category Recognition Using Classemes. ECCV (2010)
2. Li, L.J., Su, H., Lim, Y., Fei-Fei, L.: Objects as attributes for scene classification. In: (ECCV), Workshop on PaA. (2010)
3. Lazebnik, S., Schmid, C., Ponce, J.: Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In: CVPR. (2006)
4. Li, L.J., Su, H., Xing, E., Fei-Fei, L.: Object bank: A high-level image representation for scene classification & semantic feature sparsification. In: NIPS. (2010)
5. Wang, J., Yang, J., Yu, K., Lv, F., Huang, T., Gong, Y.: Locality-constrained linear coding for image classification. In: CVPR, IEEE (2010) 3360–3367
6. Yang, J., Yu, K., Gong, Y., Huang, T.: Linear spatial pyramid matching using sparse coding for image classification. In: CVPR. (2009)
7. Vogel, J., Schiele, B.: Semantic modeling of natural scenes for content-based image retrieval. IJCV **72** (2007) 133–157
8. Salakhutdinov, R., Hinton, G.: Deep Boltzmann machines. In: Proceedings of the International Conference on Artificial Intelligence and Statistics. (2009)
9. Hinton, G., Osindero, S., Teh, Y.: A fast learning algorithm for deep belief nets. Neural computation (2006)
10. Bengio, Y., Lamblin, P., Popovici, D., Larochelle, H.: Greedy layer-wise training of deep networks. (In: NIPS, 2006)
11. Olshausen, B.A., Field, D.J.: Emergence of simple-cell receptive field properties by learning a sparse code for natural images. Nature **381** (1996) 607–609
12. Grosse, R., Raina, R., Kwong, H., Ng, A.: Shift-invariant sparse coding for audio classification. In: UAI. (2007)
13. Raina, R., Battle, A., Lee, H., Packer, B., Ng, A.Y.: Self-taught learning: Transfer learning from unlabeled data. In: ICML. (2007)
14. Bengio, S., Pereira, F., Singer, Y., Strelow, D.: Group sparse coding. In: NIPS. (2009)
15. Jenatton, R., Mairal, J., Obozinski, G., Bach, F.: Proximal methods for sparse hierarchical dictionary learning. In: ICML. (2010)
16. Mairal, J., Bach, F., Ponce, J., Sapiro, G.: Online learning for matrix factorization and sparse coding. JMLR **11** (2010) 19 – 60
17. Jia, Y., Salzmann, M., Darrell, T.: Factorized latent spaces with structured sparsity. In: NIPS. (2010)
18. Olshausen, B.A., Field, D.J.: Sparse coding of sensory inputs. Current Opinion in Neurobiology **14** (2004) 481–487
19. Quattoni, A., Carreras, X., Collins, M., Darrell, T.: An efficient projection for $\ell_{1,\infty}$ regularization. In: ICML. (2009)
20. Friedman, J., Hastie, T., Tibshirani, R.: A note on the group lasso and a sparse group lasso. preprint, available at http://www-stat.stanford.edu/ tibs (2010)
21. Beck, A., Teboulle, M.: A fast iterative shrinkage-thresholding algorithm for linear inverse problems. SIAM J. Img. Sci. (2009)
22. Li, L.J., Fei-Fei, L.: What, where and who? classifying events by scene and object recognition. In: ICCV. (2007)
23. Quattoni, A., Torralba, A.: Recognizing indoor scenes. CVPR (2009)
24. Wang, C., Blei, D., Fei-Fei, L.: Simultaneous image classification and annotation. In: Proc. CVPR. (2009)