Scene Recognition on the Semantic Manifold

Roland Kwitt, Nuno Vasconcelos, and Nikhil Rasiwasia

[1] Kitware Inc., Carrboro, NC, USA
[2] Department of Electrical and Computer Engineering, UC San Diego, USA
[3] Yahoo Labs! Bangalore, India

Abstract. A new architecture, denoted *spatial pyramid matching on the semantic manifold* (SPMSM), is proposed for scene recognition. SPMSM is based on a recent image representation on a semantic probability simplex, which is now augmented with a rough encoding of spatial information. A connection between the semantic simplex and a Riemmanian manifold is established, so as to equip the architecture with a similarity measure that respects the manifold structure of the semantic space. It is then argued that the closed-form geodesic distance between two manifold points is a natural measure of similarity between images. This leads to a conditionally positive definite kernel that can be used with any SVM classifier. An approximation of the geodesic distance reveals connections to the well-known Bhattacharyya kernel, and is explored to derive an explicit feature embedding for this kernel, by simple square-rooting. This enables a low-complexity SVM implementation, using a linear SVM on the embedded features. Several experiments are reported, comparing SPMSM to state-of-the-art recognition methods. SPMSM is shown to achieve the best recognition rates in the literature for two large datasets (MIT Indoor and SUN) and rates equivalent or superior to the state-of-the-art on a number of smaller datasets. In all cases, the resulting SVM also has much smaller dimensionality and requires much fewer support vectors than previous classifiers. This guarantees much smaller complexity and suggests improved generalization beyond the datasets considered.

1 Introduction

The ability of humans to assign semantic labels (i.e., scene categories) to images, even at modest levels of attention [1], has motivated significant recent interest in image classification in computer vision (e.g., [2–7]). A popular image representation for this problem is the *bag-of-visual-features (BoF)*, an orderless collection of features extracted from the image at the nodes of an evenly-spaced grid [3]. This is used to learn a mid-level *theme* representation, which provides an image description at a higher level of abstraction. In many works [4, 8, 9], the mid-level representation consists of a codebook of *visual words*, learned in a fully unsupervised manner. The quantization of the BoF with this codebook produces a *bag-of-visual-words (BoW)* histogram, which is fed to a discriminant classifier, typically a variant of the support vector machine (SVM), for image classification. It has been shown that augmenting the BoW representation with a rough encoding of spatial information [4] and a non-linear kernel [10] can substantially boost recognition performance.

An alternative to the unsupervised theme space is to rely on *predefined* semantic themes. A set of themes is defined, a classifier trained for the detection of each theme, and each image fed to all theme classifiers. The image is finally represented by the vector of resulting classification labels. These could be binary, denoting presence/absence of the theme in the image, or graded, denoting the posterior probability of the theme given the image [11]. Since the graded representation contains all information necessary to derive the binary labels, it is the only one considered in this work. When compared to BoW, these approaches have several advantages. First, they produce a *semantic* theme space, i.e., a theme space whose coordinate axes correspond to semantic concepts. This space is usually denoted the *semantic space* (cf. [12]). It has been argued that relying on representations close to human scene understanding is as important as pure recognition accuracy [13]. Second, since the dimensionality of the semantic space is linear in the number of themes, this representation is much more compact than the high dimensional histograms required by BoW. Finally, while it has been argued that BoW lacks discriminative power [14], theme models are by definition discriminant. Hence, besides being more compact, semantic themes usually enable a more discriminative encoding of image content. When compared to BoW, the main limitation of the semantic theme representation is that theme models can lack generalization ability. This follows from the limited number of training images available per theme, much smaller than total training set size. The problem has been addressed in the literature, where different strategies have been suggested to tackle the discrimination vs. generalization trade-off, by adapting a general *background* model to the characteristics of each theme [15, 7]. A second limitation is that the theme-based representation has not been explored as extensively as the BoW. Although it could potentially benefit from the extensions developed for the latter, such as spatial information encoding and non-linear kernels, these have so far not been explored extensively. In some cases, e.g., kernel design, they are not straightforward, due to the fact that the semantic space is a *probability simplex*.

Besides classification accuracy, the computational complexity of image representations has been deemed increasingly important for image classification in the recent past. This is partly due to the emergence of large-scale benchmark datasets, such as *MIT Indoor* [5] or *SUN* [16]. In BoW methods, where recognition performance tends to increase with codebook size [17, 18], codebook generation quickly becomes a computational bottleneck. This is compounded by the need to train a kernelized classifier from a vast number of high dimensional BoW histograms. Finally, by multiplying the dimensionality of the BoW feature space by the number of spatial pyramid cells, the addition of the spatial pyramid structure of [4] can render the classification problem computationally intractable. Although the semantic space representation is much more compact than BoW, its combination with spatial encoding mechanisms and large theme vocabularies can also lead to large-scale learning problems. While in the BoW literature some authors have proposed explicit data embedding strategies [19, 20], which enable the replacement of non-linear by linear SVMs, greatly reducing computation, such embeddings are not yet available for theme-based representations.

Contribution In this work, we address several of the current limitations of the semantic theme representation by *proposing extensions of spatial information encoding, kernel design, and data embeddings compatible with image representation on a probability simplex*. This is done through the following contributions. In Section 3.1, we introduce the probability simplex as a statistical manifold and leverage principles of information geometry to derive a novel non-linear kernel on that manifold. We then adapt the spatial pyramid structure of [4] to the semantic space. Following [4], we refer to this architecture, i.e., the combination of the new kernel and the underlying semantic theme representation, as *spatial pyramid matching on the semantic manifold (SPMSM)*. In Section 3.2, we further show that the Bhattacharyya kernel is an approximation to the geodesic distance on this manifold. This leads to an explicit feature embedding, which enables the use of linear SVMs on large-scale problems. Extensive experiments, reported in Section 4, demonstrate that image classification based on the proposed SPMSM has state-of-the-art performance on a number of datasets.

2 Mid-Level Theme Representation

We start by briefly reviewing the representation of [11]. This is based on a predefined collection $\mathcal{T}$ of M themes (e.g., *sky, grass, street*). Learning is weakly supervised from a training set of images I_j, each augmented by a binary caption vector $\boldsymbol{c}^j$. Weak supervision implies that a non-zero entry at the i-th position of $\boldsymbol{c}^j$ indicates that theme i is present in image j, but a zero entry does not necessarily imply that it is absent. Images are labeled with one or more themes, which could be drawn from the set of scene category labels $\mathcal{T}$ or from another label set (e.g., scene attributes). When theme labels are the image labels, $\boldsymbol{c}^j$ contains a single non-zero entry.

As in BoW, an image I_j is represented as a collection of visual features, in some feature space $\mathcal{X}$, i.e., $I_j = \{\boldsymbol{x}_i^j\}_{i=1}^N$. These features are extracted from N localized image patches $P_i^j, \boldsymbol{x}_i^j = f(P_i^j)$. The generative model that maps an image to the semantic space is shown in the *inference* part of Fig. 1: visual features are drawn independently from themes, and themes are drawn from a multinomial random variable of parameter vector $\boldsymbol{s}^j \in [0,1]^M$. The theme occurrences of image I_j are summarized in the theme occurrence vector $(o_1^j, \ldots, o_M^j)'$. The mutinomial parameters in $\boldsymbol{s}^j$ are inferred from $\{\boldsymbol{x}_i^j\}_{i=1}^N$ as follows (the image index j is omitted for brevity). First, the theme of largest posterior probability is found per $\boldsymbol{x}_i$, i.e., $t_i^* = q_b(\boldsymbol{x}_i)$ with

$$q_b(\boldsymbol{x}_i) = \arg\max_{t \in \mathcal{T}} P_{T|\boldsymbol{X}}(t|\boldsymbol{x}_i) = \arg\max_{t \in \mathcal{T}} \frac{P_{\boldsymbol{X}|T}(\boldsymbol{x}_i|t)}{\sum_w P_{\boldsymbol{X}|T}(\boldsymbol{x}_i|w)} \quad . \tag{1}$$

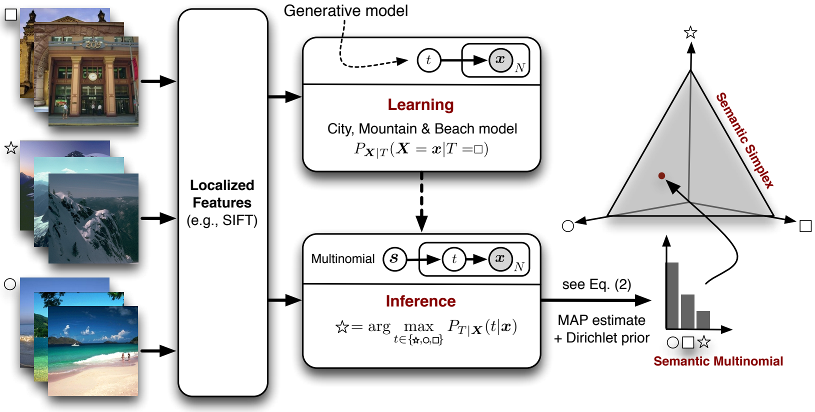
□
☆
○
Localized
Features
(e.g., SIFT)
Generative model
t
x
N
Learning
City, Mountain & Beach model
$P_{\boldsymbol{X}|T}(\boldsymbol{X}=\boldsymbol{x}|T=\square)$
Multinomial
s
t
x
N
Inference
$\star = \arg\max_{t\in\{\star,\circ,\square\}} P_{T|\boldsymbol{X}}(t|\boldsymbol{x})$
see Eq. (2)
MAP estimate
+ Dirichlet prior
○□☆
Semantic Multinomial
☆
Semantic Simplex
○
□

Fig. 1. Mapping of database images, represented by collections of visual features, to points on the semantic simplex (here $\mathbb{P}^2$).

This assumes equal prior probability for all themes, but could be easily extended for a non-uniform prior. The mapping $q_b : \mathcal{X} \to \mathcal{T}$ *quantizes* features into themes in a Bayesian, *minimum probability-of-error*, fashion. The occurrences $o_t = |\{i : t_i^* = t\}|$ of each theme t are then tallied to obtain the empirical theme occurrence vector. Finally, the MAP estimate of $\boldsymbol{s}$, for a Dirichlet prior of parameter α, is

$$\hat{\boldsymbol{s}} = \left(\frac{o_1 + \alpha - 1}{\sum_w (o_w + \alpha - 1)}, \dots, \frac{o_M + \alpha - 1}{\sum_w (o_w + \alpha - 1)} \right)' \qquad (2)$$

where α acts as a regularization parameter. In the terminology of [11], $\hat{\boldsymbol{s}}$ is denoted the *semantic multinomial (SMN)* of image I. This establishes the desired mapping $\Pi : \mathcal{X}^N \to \mathbb{P}^{M-1}$, $I \mapsto \boldsymbol{s}$ from an image represented in feature space to an image represented as a point on the *semantic (probability) simplex* $\mathbb{P}^{M-1}$.

Learning the mapping Π requires estimates of the theme-conditional distributions $P_{\boldsymbol{X}|T}(\boldsymbol{x}|t)$ from the available weakly-labeled image data. Since the theme label of *each* visual feature is not known, this is done with resort to multiple instance learning, based on the image formation model shown in the *learning* part of Fig. 1: visual features extracted from all images labeled with theme t are pooled into dataset $\mathcal{D}_t = \{\boldsymbol{x}_i^j | \boldsymbol{c}_t^j = 1\}$, which is then used to estimate $P_{\boldsymbol{X}|T}(\boldsymbol{x}|t)$. The intuition is that visual features representative of the semantic theme are more likely to occur in the training set and dominate the probability estimates. In multiple instance learning terminology, $\mathcal{D}_t$ is the *bag of positive examples* for theme t. Fig. 1 illustrates learning and inference on a three-category toy problem. Note that $\mathbb{P}^{M-1}$ serves as a new feature space for training a discriminant classifier.

3 Spatial Pyramid Matching on the Semantic Manifold

In this section we 1) introduce a statistical (semantic) manifold for image representation, 2) derive a suitable image matching kernel from the principles of information geometry and 3) augment the theme representation of the previous section with a commonly used encoding of spatial information.

3.1 The Semantic Manifold

To design a kernel for the SMN representation, one pragmatic strategy would be to choose a kernel which computes l_2 distances in feature space [18, 9], e.g., the classic RBF kernel. This, however, implicitly assumes a flat Euclidean geometry and ignores the actual geometry of the SMN data on the semantic simplex. One alternative that achieves better classification performance for BoW is the spatial pyramid match kernel (SPMK) of [10, 4], which replaces the l_2 norm by the histogram intersection (HI) metric. This, and the introduction of computationally efficient approximations [19], have made SPMK the prevalent kernel for the BoW representation.

To design a kernel suited for the SMN representation, we study the semantic simplex $\mathbb{P}^{M-1}$ in more detail. Since SMNs are parameter vectors of multinomial distributions, we equate similarity between two SMNs as the *distance* among the two associated multinomial distributions. From information geometry, it is known that $\mathbb{P}^{M-1}$ is a *Riemannian manifold*[4] if endowed with the Fisher information metric $\mathcal{I}$ (cf. [21, 22]). Hence, the distance among two SMNs $\boldsymbol{s}$ and $\boldsymbol{s}^*$ can be computed as the geodesic distance $d_{\mathcal{I}}(\boldsymbol{s}, \boldsymbol{s}^*)$ on this multinomial manifold. Although geodesics are in general hard to compute, it is possible to exploit the isomorphism $F : \mathbb{P}^{M-1} \to \mathbb{S}_+^{M-1}, \quad \boldsymbol{s} \mapsto 2\sqrt{\boldsymbol{s}}$ between the manifolds $(\mathbb{P}^{M-1}, \mathcal{I})$ and $(\mathbb{S}_+^{M-1}, \delta)$, where $\mathbb{S}_+^{M-1}$ is the positive portion of a sphere of radius two and δ denotes the Euclidean metric inherited from embedding $\mathbb{S}_+^{M-1}$ in $\mathbb{R}^M$. The isometry enables the computation of $d_{\mathcal{I}}$ as the arc on the great-circle connecting $F(\boldsymbol{s})$ and $F(\boldsymbol{s}^*)$ on the sphere, i.e.,

$$d_{\mathcal{I}}(\boldsymbol{s}, \boldsymbol{s}^*) = d_\delta(F(\boldsymbol{s}), F(\boldsymbol{s}^*)) = 2 \arccos(\langle \sqrt{\boldsymbol{s}}, \sqrt{\boldsymbol{s}^*} \rangle) \ . \qquad (3)$$

Since $\mathbb{P}^{M-1}$ is denoted the semantic simplex, we refer to $(\mathbb{P}^{M-1}, \mathcal{I})$ as the associated *semantic manifold.* It is worth mentioning that the Hellinger distance $d_H(\boldsymbol{s}, \boldsymbol{s}^*) = 2\sin(d_{\mathcal{I}}(\boldsymbol{s}, \boldsymbol{s}^*)/4)$ and the Kullback-Leibler (KL) divergence are identical to $d_{\mathcal{I}}$ up to second order as $\boldsymbol{s} \to \boldsymbol{s}^*$ [23]. The KL divergence was previously used as a similarity measure between SMNs, in a retrieval context [12], but without exploring the connections to information geometry.

These connections are particularly important for kernel design, where the metric determines the properties of the kernel. For example, the KL divergence is not symmetric and does not guarantee a positive definite kernel [24]. On

[4] A technical issue is to ensure, by (2), that SMN components are positive to guarantee that $\mathbb{P}^{M-1}$ is actually a manifold [21].

the other hand, it is known that 1) the negative of the geodesic distance $-d_{\mathcal{I}}$ satisfies all properties of a *conditionally positive definite (cpd)* kernel [22], and 2) cpd kernels can be used in any SVM classifier [25]. Consequently, we define the *semantic kernel* on the semantic manifold as

$$k(\boldsymbol{s}, \boldsymbol{s}^*) := -d_{\mathcal{I}}(\boldsymbol{s}, \boldsymbol{s}^*) \quad \boldsymbol{s}, \boldsymbol{s}^* \in \mathbb{P}^{M-1}. \tag{4}$$

As a matter of fact, the *information-diffusion kernel* of [26], specialized to the multinomial family, is an exponential (squared) variant, i.e., $\exp(-d_{\mathcal{I}}^2)$, of (4). Given a smooth-parametrization of (4), we could also leverage the work of [27], where the authors propose an adaption to SVM learning that optimizes smoothly-parametrized kernels on the simplex. While the semantic kernel might potentially benefit from those advances, we have not explored that direction in this work.

Spatial Pyramid Encoding It is now well established that augmenting the BoW representations with a rough encoding of spatial information, by means of a *spatial pyramid* [4, 28, 9], leads to significant gains in image classification. The extension of this idea to the SMN representation is quite straightforward. It suffices to compute a SMN for each of the spatial pyramid cells. Note that this introduces a *localized* semantic representation, which captures many attributes of human scene understanding. More precisely, the global SMN at pyramid level 0 captures the semantic *gist* of the image, e.g., “mostly about grass, sky, and mountains”, while SMNs at higher levels *localize* this description to each spatial pyramid cell, e.g., “mostly grass in bottom cells, mostly sky in upper cells, mostly mountains in between”. In this way, spatial cells at finer grid resolutions are more informative of local semantics and exhibit less ambiguity (cf. [13]). The structure of the SMN representation and the procedure used to estimate SMNs also enable the computation of the pyramid cell SMNs in a very efficient manner. In fact, it suffices to compute the SMNs of the pyramid cells at the finest grid resolution. The SMN of index n at the overlying pyramid level l can then be directly inferred from its four child-cell SMNs $\{\boldsymbol{s}_{l+1,4n+i}\}_{i=0}^{3}$, at level $l+1$, by computing the convex combination $\boldsymbol{s}_{l,n} = {}^1\!/\!_4 \cdot (\boldsymbol{s}_{l+1,4n} + \cdots + \boldsymbol{s}_{l+1,4n+3})$. In other words, the SMN of one spatial pyramid cell at level l lies in the *convex hull* spanned by its four child-cell SMNs at the next finer level. In total, there are ${}^1\!/\!_3 \cdot (4^L - 1)$ SMNs per image, for a spatial pyramid with L levels.

In order to incorporate spatial constraints in the classification, it is possible to combine the semantic kernel with the spatial pyramid structure, in a way similar to [4]. This consists of 1) assigning more weight to matches at finer pyramid resolutions and 2) normalizing the geodesic distances at one pyramid level by the number of grid cells at that level. Given two images I_a and I_b, represented by their concatenated SMNs $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, the *semantic spatial pyramid match kernel (SSPMK)* is defined as

$$k(\boldsymbol{\alpha}, \boldsymbol{\beta}) = -\sum_{l=0}^{L-1} w_l \sum_{n=0}^{4^l} d_{\mathcal{I}}(\varphi_{l,n}(\boldsymbol{\alpha}), \varphi_{l,n}(\boldsymbol{\beta})) \tag{5}$$

with $w_l := \tilde{w}_l \bar{w}_l$, where $\tilde{w}_l = {}^1/_{4^l}$ denotes the *normalization* weight at level l and $\bar{w}_l = 2^{-(L+l)}$ denotes the corresponding *matching* weight. Note that we used $\varphi_{l,n}(\boldsymbol{\alpha}) = \boldsymbol{s}_{l,n}$ to denote the extraction of $\boldsymbol{s}_{l,n}$ from a concatenated SMN vector $\boldsymbol{\alpha}$. Since (5) is a weighted sum of semantic kernels, and the closure property for weighted sums of positive definite kernels extends to the family of cpd kernels [25], the SSPMK is a cpd kernel.

3.2 Data Embedding

Given the SMN representation, it remains to train an SVM classifier. For small-scale datasets, it is feasible to learn a non-linear SVM, albeit the training complexity is somewhere between quadratic and cubic [20]. In general, however, non-linear SVMs do not scale well with training set size. On large-scale problems, linear SVMs are overwhelmingly preferred due to their efficient (i.e., linear-time) training algorithms. The question is how to rely on a linear SVM, but still somehow exploit the power of the SSPMK. Ideally, it would be possible to derive an explicit SMN embedding that preserves the advantages of the geodesic distance. The training of a non-linear SVM for SMN classification could then be reduced to training a linear SVM on the embedded features. Unfortunately, exact embeddings are rarely available. Although approximations are possible, these usually entail a loss in recognition performance.

While a popular embedding exists for the HI kernel [19], it exploits the additivity property of the kernel. Since the semantic kernel of (4) is not additive, neither the embedding of [19], nor the embedding learning method of [20] are feasible. One alternative, that we explore, is to replace the arccos term by a first-order Taylor series around 0, i.e., $\arccos(x) \approx {}^\pi/_2 - x + \mathcal{O}(x)^2$. This leads to the approximation of (4) by

$$k(\boldsymbol{s}, \boldsymbol{s}^*) \approx -\pi + \langle \sqrt{\boldsymbol{s}}, \sqrt{\boldsymbol{s}^*} \rangle \quad \boldsymbol{s}, \boldsymbol{s}^* \in \mathbb{P}^{M-1} \ . \tag{6}$$

Notably, the dot-product on the right-hand side is the additive Bhattacharyya kernel of [20]. Although, a linear approximation can be coarse, the ability to immediately read the explicit data embedding $\phi(x) = \sqrt{x}$ is appealing, since it entails almost no computational cost. Finally, taking the spatial pyramid structure into account, the extended embedding for the n-th SMN at pyramid level l can be written as

$$\phi(\boldsymbol{s}_{n,l}) = \sqrt{w_l \boldsymbol{s}_{n,l}} \ . \tag{7}$$

While the Bhattacharyya kernel has previously been used in image classification (cf. [29]), its practical success now has another principled justification due to the close relationship with the geodesic distance.

4 Experiments

In this section, we report on a number of experiments designed to evaluate the classification accuracy of the proposed SPMSM architecture, i.e., the combination of the SMN representation of (2) and the SSMPK of (5).

4.1 Datasets & Implementation

Three popular, yet rather small, benchmark datasets and two recent mid- to large-scale datasets were used in our recognition experiments. The smaller ones are the *LabelMe* [2], UIUC *Sports* [30] and *15 Scenes (N15)* [3,4] datasets. For mid- to large-scale experiments, we used the *MIT Indoor* [5] scenes and the *SUN* [16] dataset. We use the prevalent training/testing configurations in the literature. Recognition rates on *LabelMe*, *Sports* and *N15* were averaged over three test runs with random training/testing splits. In the case of *MIT Indoor* and *SUN*, the training/testing configurations are provided by the original authors. All images were converted to grayscale and resized to have maximum dimension of 256 pixels (while maintaining the aspect ratio).

The appearance representation was based on SIFT[5] descriptors [31], computed on an evenly-spaced 4×4 pixel grid. 128-component Gaussian mixtures of diagonal covariance were used to model theme distributions, and mixture parameters estimated with the EM algorithm (initialized by K-Means++). We chose a directed mixture parameter estimation approach in contrast to the hierarchical estimation procedure employed in [12,11]. All experiments involving spatial pyramids relied on three pyramid levels. Further refinements did not produce improvements, confirming the findings of [4]. For the tests on *LabelMe*, *Sports*, *N15* and *MIT Indoor*, we used the LIBSVM [32] implementation of a C-SVM and a 1-vs-1 multi-class classification strategy. On feature embedding experiments, we relied on LIBLINEAR [33] to train a linear SVM and switched from 1-vs-1 to 1-vs-all multi-class classification, for performance reasons. The SVM cost factor C was determined by three-fold cross-validation on the training data, evaluated at 20 linearly spaced positions of $\log C \in [-2, 4]$.

4.2 Evaluation

Semantic Kernel The first set of experiments was designed to evaluate the semantic kernel of (4). In all cases, the image representation was the SMN of (2). We started with a comparison to two popular kernels in the literature: HI (k_{HI}), and χ^2 (k_{χ^2}). The kernel definitions are given in Table 1 for two input vectors $\boldsymbol{x}, \boldsymbol{y} \in [0,1]^M$. It is worth noting that SVM training with one of these kernels only requires tuning of the cost factor C, whereas RBF variants require tuning of the kernel width as well. The table presents the recognition accuracies obtained on *Sports*, *LabelMe* and *N15.* The semantic kernel achieves the highest average rate on all datasets. This illustrates the benefits of adopting a kernel which is tailored to the manifold structure of the semantic space.

Spatial Pyramid Encoding We next considered the full SPMSM architecture, by augmenting the semantic kernel with SPM. This was compared to the standard implementation of SPM with the kernels of the previous experiment. Results are listed in Table 1. Two conclusions are possible from the table. First,

[5] LEAR impl.: `http://lear.inrialpes.fr/people/dorko/downloads.html`

Table 1. Comparison of the semantic kernel to the HI (k_{HI}) and the χ^2 kernel (k_{χ^2}) *without* and *with* SPM.

Kernel Type	**without** SPM			**with** SPM		
	Sports	*LabelMe*	*N15*	*Sports*	*Labelme*	*N15*
Proposed, see (4), (5)	**79.1**	**84.7**	**79.1**	**83.0**	**87.5**	**82.3**
$k_{\chi^2}, \sum_i \frac{x_i y_i}{(x_i+y_i)}$	78.6	84.6	78.9	81.6	86.2	81.0
$k_{HI}, \sum_i \min(x_i, y_i)$	77.8	84.1	78.6	81.8	87.0	82.0

the addition of the spatial pyramid structure does not change the relative performances of the kernels: the gap in recognition performance between SPM with k_{HI} and SPMSM is similar to that between the HI (k_{HI}) and the semantic kernel when omitting SPM. Second, the results are consistent with previous reports on the benefits of spatial information encoding [4]. Comparing the results *with* and *without* SPM shows that, for the SSPMK, this gain is around three to four percentage points. In addition, we remark that training with a RBF kernel, optimizing the cost factor and kernel width on a 2-D grid, exhibits performance similar to the worst result per kernel on each database of Table 1 (with and without SPM). This underpins the assertion (cf. [34, 27]) that kernels which are effective in Euclidean space (like RBF) are not necessarily effective in another space, such as the semantic manifold.

Data Embedding Finally, we evaluated the semantic kernel approximation of Section 3.2 and the square-root embedding of (6). This was compared to the popular HI kernel embedding of [19] and to a linear SVM *without* any embedding, i.e., applied directly to the SMNs. The comparison to [19] was performed against the sparse ϕ_2 embedding[6] (denoted as ϕ_2^s in the original work) with ten discrete levels. Table 2 lists the recognition rates on all datasets, without spatial pyramid matching. A few conclusions are possible from the table. First, the advantages of using the kernel+embedding combination are not very significant for small datasets. In fact, [19] underperformed the linear SVM without embedding on semantic space, on all three small datasets. While the square-root embedding outperformed the latter, the gains were relatively small. Second, a different picture emerges for the large datasets, where both embeddings outperformed the SVM without embedding. Again, the square-root embedding achieved the best performance, now with non-trivial gains over the two other approaches. Third, the square-root embedding outperformed the embedding of [19], preserving the advantages of the semantic kernel on all datasets. Finally, although there is a drop in recognition rate when compared to Table 1, this drop is small (about one to two percentage points). We believe that the computational savings associated with a linear SVM far outweigh this slight loss in recognition performance.

[6] available from `http://www.cs.berkeley.edu/~smaji/projects/add-models/`

Table 2. Comparison (*without* SPM) of the proposed feature embedding to that of [19] and no embedding.

Dataset	**Embedding Variant**		
	Maji & Berg [19]	Proposed	Without
Sports	76.9	**77.8**	77.1
LabelMe	83.0	**84.3**	84.0
N15	76.8	**77.3**	77.0
MIT Indoor	32.2	**33.7**	31.9
SUN	23.1	**24.3**	22.0

Comparing to Bag-of-Words This set of experiments was designed to compare SPMSM to the combination of BoW and SPM, which can be considered a de-facto standard for image classification. However, the comparison turned out not to be straightforward. For example, it is well known that the performance of BoW methods increases with codebook size. This is, in significant part, due to the associated increase in the dimensionality of the SVM that ultimately classifies the images. In general, the performance of an SVM improves with the dimensionality of its input, as long as the latter remains in a reasonable range. The problem is that SPMSM and BoW+SPM can have very different SVM dimensionalities.

Without the spatial pyramid structure, this dimensionality equals the number of themes, for SMN, and the number of codewords, for BoW. With the spatial pyramid, these numbers are multiplied by the number of spatial pyramid cells, which is 21 for three pyramid levels. Since there are as many themes as scene category labels, SPMSM has a fixed SVM dimensionality. On the other hand, it is always possible to increase the codebook size of BoW. While this suggests using SPMSM as a reference, its dimensionality is usually too low for BoW+SPM, which performs quite poorly for codebook cardinalities equivalent to the number of scene categories. An alternative would be to increase the dimensionality of SPMSM, e.g., by replacing the hard assignment of (1) with a histogram of the posterior probabilities $P_{T|\boldsymbol{X}}(t|\boldsymbol{x}_i)$ for each theme t.

We have not considered such possibilities, simply measuring the recognition rate of BoW+SPM for various values of the codebook size. The recognition rates are shown in Table 3. Rates higher than those achieved by SPMSM are marked in bold, whereas rates at *equivalent dimensionality* are underlined. It is clear that BoW+SPM requires a much higher dimensionality than SPMSM, for equivalent performance. For the datasets considered, the ratio of dimensionalities is ≈ 30. While this may not be a problem for the small corpora that are commonly used in the literature, e.g., the eight category *LabelMe* or *Sports* datatsets, it can be much more problematic for richer corpora, such as *MIT Indoor* or *SUN*. Even on the modestly sized *N15* dataset, SPM+BoW requires a codebook of size 512 to guarantee a minor gain over SPMSM. This corresponds to a SVM of $512 \times 21 = 10,752$ dimensional input, as opposed to the $15 \times 21 = 315$ dimensions of SPMSM. From the trend in Table 3, the aforementioned threshold would likely

Table 3. Recognition rate of BoW+SPM, for varying codebook sizes. Results higher than those achieved by SPMSM (shown at the bottom) are marked bold; results at *equal SVM dimensionality* are underlined. Numbers in parentheses denote the percentage of the training examples selected as support vectors.

Codebook	**Dataset**			
	LabelMe	*Sports*	*N15*	*MIT Indoor*
8	74.0 (74)	64.8 (87)	63.0 (89)	19.1 (98)
16	78.8 (75)	69.3 (88)	69.7 (89)	25.3 (98)
32	82.9 (75)	77.7 (88)	73.6 (89)	32.6 (98)
64	85.9 (75)	80.4 (89)	77.9 (89)	36.2 (99)
128	87.4 (77)	81.4 (90)	80.8 (90)	38.8 (99)
256	**88.0** (79)	**83.6** (91)	81.7 (91)	41.0 (99)
512	**88.6** (82)	**84.7** (92)	**83.1** (93)	43.6 (99)
SPMSM	87.5(57)	83.0 (67)	82.3 (74)	**44.0**(95)

occur at 43,008 dimensions for *MIT Indoor*. Since this exceeds the capacity of the SVM package that we have used in these experiments we could not even confirm if BoW+SPM can actually outperform SPMSM (dimensionality 1,407) on this dataset.

Another factor that confounds the comparison of the two approaches is the type of support vectors that they produce. In fact, the percentage of examples that an SVM chooses as support vectors is a well known measure of the difficulty of the classification, and the degree to which the classifier is "overfitting to the dataset", i.e., modeling the intricacies of the particular dataset where performance is evaluated, rather than learning a truly generic decision rule. The numbers in parenthesis in Table 3 show the support vector percentages of BoW+SPM, for various codebook sizes, and SPMSM. Note that the percentages are indeed higher for the datasets of lower recognition rate. It is also clear that, on the harder datasets, the BoW+SPM SVM considers virtually every training example a support vector. The fact that the SPMSM SVM achieves near equivalent recognition rates with much smaller support vector percentages indicates that the classification is much *easier* on the semantic manifold. Hence, SPMSM is likely to generalize much better if applied to data collected from other sources.

In summary, on the large datasets considered, SPMSM has state-of-the-art performance. On the remaining, its performance is superior to that of BoW+SPM, by a large margin, for SVMs of equivalent dimensionality. On all datasets, it took BoW+SPM a 30-fold increase in dimensionality to achieve results similar to those of SPMSM, if at all. The percentages of examples selected as support vectors also suggest that classification is much simpler on the semantic manifold, and that SPMSM is likely to generalize better to unseen datasets. Computationally, since SVM complexity is linear on the *product* of the number of support vectors and dimensionality, the SPMSM SVM is significantly less challenging to implement.

Table 4. Comparison to the state-of-the-art.

Dataset	State-of-the-Art	Rate [%]
Sports	Li & Fei-Fei [30]	73.4
	Proposed	83.0
	Wu & Rehg [28]	**84.3**
LabelMe	Wang et al. [35]	76.0
	Dixit et al. [7]	86.9
	Proposed	**87.5**
N15	Lazebnik et al. [4]	81.2
	Proposed	82.3
	Dixit et al. [7]	**85.4**
MIT Indoor	Quattoni & Torralba [5]	25.0
	Pandey & Lazebnik [36]	43.1
	Proposed	**44.0**
SUN	Xiao et al. [16]	27.2
	Proposed	**28.9**

Comparing to the State-of-the-Art Finally, we compare SPMSM to the state-of-the-art in the literature. An overview of the recognition rates of various methods is given in Table 4. Note that a direct comparison of the different methods is not totally fair, since they differ along many dimensions, not just the kernel. In fact, many of the BoW enhancements at the core of these methods could be applied to the SPMSM. Nevertheless, the results of SPMSM classification are excellent: to the best of our knowledge, the proposed classifier has the highest published rates on the large- and mid-scale datasets (*SUN* and *MIT Indoor*), and one of the small-scale ones (*LabelMe*). On *MIT Indoor*, it substantially outperforms the baseline of [5], and does slightly better than the previous best approach of [36]. A detailed comparison to [5] is shown in Fig. 2. The improvements are distributed across all indoor scene categories: there are only 11 classes where SPMSM performs at a similar or worse level. With respect to the remaining datasets, *Sports* and *N15*, SPMSM outperforms the baseline and achieves results competitive with the best in both cases. Note, for example, that the best method on *N15* [7] specifically addresses the generalization ability of theme models, through model adaptation techniques. Since these techniques could equally be used to improve the theme models of SPMSM, the two methods are *complementary,* not competitors. We plan to include model adaptation in SPMSM in future work.

Acknowledgements This work was supported, in part, by the NIH/NIBIB grant "National Alliance of Medical Image Computing" (1U54EB005149), the NIH/NCI grant "Image Registration for Ultrasound-Based Neurosurgical Navigation" (1R01CA138419) and the NSF award CCF-0830535.

dentaloffice	42.9	57.1	**(14.2)**
stairscase	30.0	35.0	**(5)**
children_room	5.6	44.4	**(38.8)**
hospital_room	35.0	35.0	(0)
closet	38.9	77.8	**(38.9)**
bar	22.2	38.9	**(16.7)**
warehouse	9.5	33.3	**(23.8)**
grocerystore	38.1	42.9	**(4.8)**
buffet	55.0	65.0	**(10)**
classroom	50.0	50.0	(0)
inside_subway	23.8	71.4	**(47.6)**
corridor	38.1	57.1	**(19)**
jewelleryshop	0.0	27.3	**(27.3)**
prisoncell	10.0	45.0	**(35)**
operating_room	10.5	31.6	**(21.1)**
pool_inside	25.0	50.0	**(25)**
hairsalon	9.5	33.3	**(23.8)**
locker_room	38.1	38.1	(0)
elevator	61.9	66.7	**(4.8)**
concert_hall	45.0	55.0	**(10)**
restaurant_kitchen	4.3	30.4	**(26.1)**
gameroom	25.0	30.0	**(5)**
bookstore	20.0	55.0	**(35)**
inside_bus	39.1	60.9	**(21.8)**
auditorium	55.6	55.6	(0)
kindergarden	5.0	55.0	**(50)**
lobby	10.0	25.0	**(15)**
deli	21.1	15.8	(-5.3)
computerroom	44.4	50.0	**(5.6)**
videostore	27.3	36.4	**(9.1)**
movietheater	15.0	45.0	**(30)**
trainstation	35.0	75.0	**(40)**
museum	4.3	21.7	**(17.4)**
clothingstore	22.2	44.4	**(22.2)**
mall	0.0	20.0	**(20)**
kitchen	23.8	47.6	**(23.8)**
dining_room	16.7	27.8	**(11.1)**
bathroom	33.3	33.3	(0)
church_inside	63.2	68.4	**(5.2)**
meeting_room	9.1	31.8	**(22.7)**
restaurant	5.0	30.0	**(25)**
nursery	35.0	47.4	**(12.4)**
toystore	13.6	40.9	**(27.3)**
shoeshop	5.3	21.1	**(15.8)**
pantry	25.0	65.0	**(40)**
livingroom	15.0	10.0	(-5)
bowling	45.0	75.0	**(30)**
tv_studio	27.8	50.0	**(22.2)**
library	40.0	50.0	**(10)**
bakery	15.8	31.6	**(15.8)**
studiomusic	36.8	36.8	(0)
florist	36.8	73.7	**(36.9)**
gym	27.8	22.2	(-5.6)
cloister	45.0	80.0	**(35)**
greenhouse	50.0	70.0	**(20)**
waitingroom	19.0	19.0	(0)
bedroom	14.3	47.6	**(33.3)**
laboratorywet	0.0	40.9	**(40.9)**
winecellar	23.8	28.6	**(4.8)**
casino	21.1	57.9	**(36.8)**
office	0.0	38.1	**(38.1)**
fastfood_restaurant	23.5	64.7	**(41.2)**
airport_inside	10.0	10.0	(0)
laundromat	31.8	40.9	**(9.1)**
artstudio	10.0	25.0	**(15)**
subway	9.5	42.9	**(33.4)**
garage	27.8	55.6	**(27.8)**

Fig. 2. Detailed comparison of the recognition performance of SPMSM and the baseline of [5] on *MIT Indoor*. The difference is given in parentheses. Scenes where SPMSM performs worse are marked red (best viewed in color).

References

1. Fei-Fei, L., VanRullen, R., Koch, C., Perona, P.: Rapid natural scene categorization in the near absence of attention. PNAS **99** (1999) 9566 – 9601
2. Olivia, A., Torralba, A.: Modeling the shape of a scene: A holistic representation of the spatial envelope. IJCV **42** (2001) 145 – 175,
3. Fei-Fei, L., Perona, P.: A bayesian hierarchical model for learning natural scene categories. In: CVPR. (2005)
4. Lazebnik, S., Schmid, C., Ponce, J.: Beyond bags of features: Spatial pyramid matching for recognizing scene categories. In: CVPR. (2006)
5. Quattoni, A., Torralba, A.: Recognizing indoor scenes. In: CVPR. (2009)
6. 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)
7. Dixit, M., Rasiwasia, N., Vasconcelos, N.: Adapted gaussian mixtures for image classification. In: CVPR. (2011)
8. Bosch, A., Zisserman, A., Munoz, X.: Image classification with random forests and ferns. In: ICCV. (2007)
9. Wu, J., Rehg, J.: CENTRIST: A visual descriptor for scene categorization. PAMI **33** (2011) 1489–1501
10. Grauman, K., Darrell, T.: Pyramid match kernels: Discriminative classification with sets of image features. In: ICCV. (2005)
11. Rasiwasia, N., Vasconcelos, N.: Scene classification with low-dimensional semantic spaces and weak supervision. In: CVPR. (2008)
12. Rasiwasia, N., Moreno, P., Vasconcelos, N.: Bridging the gap: Query by semantic example. IEEE Trans. Multimedia **9** (2007) 923–938

13. Schwaninger, A., Vogel, J., Hofer, F., Schiele, B.: A psychophysically plausible model for typicality ranking of natural scenes. ACM Trans. Appl. Percept. **3** (2006) 333–353
14. Boiman, O., Shechtman, E., Irani, M.: In defense of nearest-neighbor based image classification. In: CVPR. (2008)
15. Perronnin, F.: Universal and adapted vocabularies for generic visual categorization. PAMI **30** (2008) 1243–1256
16. Xiao, J., Hayes, J., Ehringer, K., Olivia, A., Torralba, A.: SUN database: Large-scale scene recognition from Abbey to Zoo. In: CVPR. (2010)
17. Nowak, E., Jurie, F., Triggs, B.: Sampling strategies for bag-of-features image classification. In: ECCV. (2006)
18. Csurka, G., Dance, C., Fan, L., Willamowski, J., Bray, C.: Visual categorization with bags of keypoints. In: Proceedings of the International Workshop on Statistical Learning in Computer Vision. (2004)
19. Maji, S., Berg, A.: Max-margin additive classifiers for detection. In: ICCV. (2009)
20. Perronnin, F.: Large-scale image categorization with explicit data embedding. In: CVPR. (2010)
21. Lebanon, G.: Riemannian Geometry and Statistical Machine Learning. PhD thesis, Carnegie Mellon University (2005)
22. Zhang, D., Chen, X., Lee, W.: Text classification with kernels on the multinomial manifold. In: ACM SIGIR. (2005)
23. Kaas, R.: The geometry of asymptotic inference. Stat. Sci. **4** (1989) 188–219
24. Moreno, P., Ho, P., Vasconcelos, N.: A Kullback-Leibler divergence based kernel for SVM classification in multimedia applications. In: NIPS. (2003)
25. Schölkopf, B., Smola, A.: Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond. MIT Press, Cambridge, MA, USA (2001)
26. Lafferty, J., Lebanon, G.: Diffusion kernels on statistical manifolds. JMLR **6** (2005) 129 – 163
27. Ablavsky, V., Sclaroff, S.: Learning parameterized histogram kernels on the simplex manifold for image and action classification. In: ICCV. (2011)
28. Wu, J., Rehg, J.: Beyond the euclidean distance: Creating effective visual codebooks using the histogram intersection kernel. In: ICCV. (2009)
29. Zhou, X., Yu, K., Zhang, T., Huang, T.: Image classification using super-vector coding of local image descriptors. In: ECCV. (2010)
30. Li, L.J., Fei-Fei, L.: What, where and who? classifying events by scene and object recognition. In: ICCV. (2007)
31. Lowe, D.: Distinctive image features from scale-invariant keypoints. IJCV **60** (2004) 91–110
32. Chang, C.C., Lin, C.J.: LIBSVM: A library for support vector machines. ACM TIST **2** (2011) 1–27
33. Fan, R.E., Chang, K.W., Hsieh, C.J., Wang, X.R., Lin, C.J.: LIBLINEAR: A library for large linear classification. JMLR **9** (2008) 1871–1874
34. Chapelle, O., Haffner, P., Vapnik, V.: Support vector machines for histogram-based image classification. IEEE Trans. Neural Netw. **10** (1999) 1055 – 1064
35. Wang, C., Blei, D., Fei-Fei, L.: Simultaneous image classification and annotation. In: CVPR. (2009)
36. Pandey, M., Lazebnik, S.: Scene recognition and weakly supervised object localization with deformable part-based models. In: ICCV. (2011)