Appears in IEEE Conf. on Computer Vision and Pattern Recognition, Anchorage, 2008.

Scene Classification with Low-dimensional Semantic Spaces and Weak Supervision

Nikhil Rasiwasia Nuno Vasconcelos
Department of Electrical and Computer Engineering
University of California, San Diego
nikux@ucsd.edu, nuno@ece.ucsd.edu

Abstract

A novel approach to scene categorization is proposed. Similar to previous works of [11, 15, 3, 12], we introduce an intermediate space, based on a low dimensional semantic "theme" image representation. However, instead of learning the themes in an unsupervised manner, they are learned with weak supervision, from casual image annotations. Each theme induces a probability density on the space of low-level features, and images are represented as vectors of posterior theme probabilities. This enables an image to be associated with multiple themes, even when there are no multiple associations in the training labels. An implementation is presented and compared to various existing algorithms, on benchmark datasets. It is shown that the proposed low dimensional representation correlates well with human scene understanding, and is able to learn theme co-occurrences without explicit training. It is also shown to outperform unsupervised latent-space methods, with much smaller training complexity, and to achieve performance close to the state of the art methods, which rely on much higher-dimensional image representations. Finally a study of the effect of dimensionality on the classification performance is presented, indicating that the dimensionality of theme space grows sub-linearly with the number of scene categories.

1. Introduction

Scene classification is an important problem for computer vision, and has received considerable attention in the recent past. It differs from the conventional object detection/classification, to the extent that a scene is composed of several entities often organized in an unpredictable layout [15]. For a given scene, it is virtually impossible to define a set of properties that would be inclusive of all its possible visual manifestations. Frequently, images from two different scene categories are visually similar, e.g., it can be difficult to distinguish between scenes such as "Street" and "City" (see Sec. 4).

Early efforts at scene classification targeted binary problems, such as distinguishing indoor from outdoor scenes [18] etc. Subsequent research was inspired by the literature on human perception. In [1], it was shown that humans can recognize scenes by considering them in a "holistic" manner, without recognizing individual objects. Drawing inspiration from the perceptual literature, [14] proposed a low dimensional representation of scenes, based on several global properties such as "naturalness", "openness", etc. More recently, there has been an effort to solve the problem in greater generality, through design of techniques capable of classifying relatively large number of scene categories [20, 11, 15, 10, 3, 12]. These methods tend to rely on *local region descriptors*, modeling an image as an order less collection of descriptors, commonly known as the "bag-of-features". The space of local region descriptors is then quantized, based on some clustering mechanism, and the mean vectors of these clusters, commonly known as "visterms"[1] are chosen as their representatives. The representation of an image in this quantized space, is referred to as the "bag-of-visterms" representation. A set of cluster means, or visterms, forms a "codebook", and a scene is characterized as a frequency vector over the visterms in the codebook [5]. This representation is motivated by the time-tested "bag-of-words" model, widely used in text-retrieval [16]. The analogy between visual-words and text-words is also explored in [17].

Lately, various extensions of this basic "bag-of-visterms" model have been proposed [11, 15, 3, 12]. All such methods aim to provide a compact lower dimensional representation using some intermediate characterization on a latent space, commonly known as the intermediate "theme" or "topic" representation [11]. The rationale is that images which share frequently co-occurring visterms have similar representation in the latent space, even if they have

[1]In the literature the terms "textons", "keypoints", "visual-words", "visual-terms" or "visterms" have been used with approximately the same meaning, i.e. mean vectors of the clusters in a high-dimensional space.

no visterms in common. This leads to representations robust to the problems of polysemy - a single visterm may represent different scene content, and synonymy - different visterms may represent the same content [15]. It also helps to remove the redundancy that may be present in the basic "bag-of-visterms" model, and provides a semantically more meaningful image representation. Moreover, a lower dimensional latent space speeds up computation: for example, the time complexity of a Support Vector Machine (SVM) is linear in the dimension of the feature space. Finally, it is unclear that the success of the basic "bag-of-visterms" model would scale to very large problems, containing both large image corpuses and a large number of scene categories. In fact, this has been shown not to be the case in text-retrieval, where it is now well established that a flat representation is insufficient for large scale systems, and the use of intermediate latent spaces leads to more robust solutions [8, 2]. However, a direct translation of these methods to computer vision has always incurred a loss in performance, and latent models have not yet been shown to be competitive with the flat "bag-of-visterms" representation [12, 10].

In this paper we propose an alternative solution. Like the latent model approaches, we introduce an intermediate space, based on a low dimensional semantic "theme" representation. However, instead of learning the themes in an unsupervised manner, from the "bag-of-visterms" representation, the semantic themes are explicitly defined, and the images are casually annotated with respect to their presence[2]. This can *always* be done since, in the absence of "thematic" annotations, the "themes" can be made equal to the class labels, which are always available. The number of semantic themes used defines the dimensionality of the intermediate theme space, henceforth referred to as "semantic space". Each theme induces a probability density on the space of low-level features, and the image is represented as the vector of posterior theme probabilities. An implementation of this approach is presented and compared to existing algorithms on benchmark datasets. It is shown that the proposed low dimensional representation correlates well with human scene understanding, captures theme co-occurrences without explicit training, outperforms the unsupervised latent-space approaches, and achieves performance close to the state of the art, previously only accessible with the flat "bag-of-visterms" representation, using a much higher dimensional image representation.

2. Related Work

Low dimensional representations for scene classification have been studied in [11, 15, 3, 12]. On one hand, it is noticed that increasing the size of the codebook improves classification performance[13]. Csurka et al. [5] compare different codebook sizes ranging from 100 to 2500 visterms, showing that performance degrades monotonically as size decreases. Quelhas et al. [15] also experience a monotonic degradation of performance for 3-class classification, and use a codebook of 1000 visterms. In [10], Lazebnik et al. show that performance increases when codebook size is increased from 200 to 400 visterms.

On the other hand, there is a strong desire for low dimensional representations, for the benefits elucidated in Sec. 1. This is achieved by resorting to techniques from the text-processing literature, such as Latent Dirichlet Allocation (LDA) [2], Probabilistic Latent Semantic Analysis (pLSA) [8] etc, which produce an intermediate latent "theme" representation. Fei-Fei et al. [11] motivate the use of intermediate representations, citing the use of "textons" in texture retrieval. They then propose two variations of LDA to generate the intermediate theme representation. In [15], Quelhas et al. use pLSA, to generate the compact representation. They argue that pLSA has the dual ability to generate a robust, low dimensional scene representation, and to automatically capture meaningful scene aspects or themes. pLSA is also used by Bosch et al. in [3]. Another approach to two-level representation based on the Maximization of Mutual Information (MMI) is presented in [12]. However, a steep drop in classification performance is often experienced as a result of dimensionality reduction [12, 10].

3. Proposed Approach

A scene classification system can be broadly divided into two modules. The first defines the image representation, while the second delineates the classifier used for decision making. Since the main goal of this work is to present a low-dimensional semantic theme representation, we do not duel on the choice of classifier, simply using an SVM. This is the standard choice in the scene classification literature [20, 13, 5]. To obtain the semantic theme representation, the image is first represented as a bag of localized descriptors on a space of low-level features, which is then mapped to the space of semantic themes using machine learning techniques. This is similar in principle to the two level image representations of [11, 15, 3, 12].

3.1. Image Representation

We start by formalizing the image representation.

3.1.1 Low-level Representation

Consider a labeled image database $\mathcal{D} = \{(\mathcal{I}_1, \mathbf{s}_1), \ldots, (\mathcal{I}_D, \mathbf{s}_d)\}$ where images $\mathcal{I}_i$ are observations from a random variable $\mathbf{X}$, defined on some feature space $\mathcal{X}$. Each image is represented as a set of n *low-level feature vectors* $\mathcal{I} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}, \mathbf{x}_i \in \mathcal{X}$, which are vectors of *localized descriptors*, assumed to be sampled

[2]Here, "casually" means that the image may only be annotated with a subset of the themes that it actually contains.

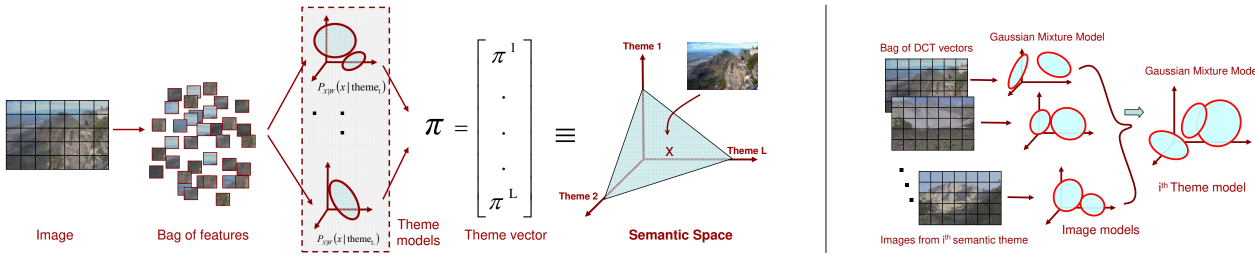


Figure 1. Left) The proposed scene classification architecture. Right) Learning the semantic theme density from the set $\mathcal{D}_t$ of all training images annotated with the t^{th} caption in $\mathcal{L}$, using hierarchical estimation [4].

independently. This framework is common to the "bag-of-visterms" representation, where each local descriptor $\mathbf{x}_i$ is further quantized into one of the visterms according to a nearest neighbor rule [15]. Here, however, we do not rely on such quantization. The image labels $\mathbf{s}_i$ is considered to be an observation from a semantic scene category S defined on $\{s_1, \ldots, s_K\}$. Note that the label $\mathbf{s}_i$ is an indicator vector such that $\mathbf{s}_{i,j} = 1$ if the i^{th} image is an observation from the j^{th} scene category.

3.1.2 Semantic Theme Representation

To represent images by semantic themes, the database $\mathcal{D}$ is augmented with a vocabulary $\mathcal{L} = \{t_1, \ldots, t_L\}$ of semantic themes t_i, and each image $\mathcal{I}_i$ with a pre-specified caption $\mathbf{c}_i$, making $\mathcal{D} = \{(\mathcal{I}_1, \mathbf{s}_1, \mathbf{c}_1), \ldots, (\mathcal{I}_D, \mathbf{s}_D, \mathbf{c}_D)\}$. Here $\mathbf{c}_i$ is a binary L-dimensional vector such that $\mathbf{c}_{i,j} = 1$ if the i^{th} image was annotated with the j^{th} theme in $\mathcal{L}$. Themes are drawn from a random variable T, which takes values in $\{t_1, \ldots, t_L\}$. Each theme induces a probability density $\{P_{\mathbf{X}|T}(\mathbf{x}|t_i)\}_{i=1}^{L}$ on $\mathcal{X}$, from which feature vectors are drawn. In general, themes are different from image classes. For example, images in the "Street" class of Figure 2 contain themes such as "road", "sky", "people", or "cars". However, in the absence of "theme" annotations in the training dataset, the set of semantic scene categories $\{s_1, \ldots, s_K\}$, e.g. "Street", can serve as a proxy for the theme vocabulary. In this case, each image is only explicitly annotated with one "theme", even though it may depict multiple: e.g. most images in the "Street" class of Fig. 2 also depict "Buildings". We refer to this limited type of scene labeling as *casual annotation*. This is the annotation mode for all results reported in this paper, to enable comparison to previous scene classification work. We will see that supervised learning of the intermediate theme space with casual annotations can be far superior to unsupervised learning of a latent theme space, as previously proposed [11].

3.1.3 Scene Classification

Due to the limited information contained in casual annotations, images cannot be simply represented by the caption vectors $\mathbf{c}_i$. In fact, $\mathbf{c}_i$ is only available for training images, and $\mathbf{c}_{i,j} = 0$ does not mean that the i^{th} image does not contain the j^{th} theme, simply that it was not annotated with it. Instead, the proposed classification system represents images by vectors of theme frequency, or counts, $\mathcal{I} = (f_1, \ldots, f_L)^T$. Each low level feature vector extracted from an image is assumed to be sampled independently from the probability distribution of a semantic theme, and f_i is the number of vectors drawn from the i^{th} theme. In this way, an image can be associated with multiple themes, even when there are no multiple associations in the labels used for training.

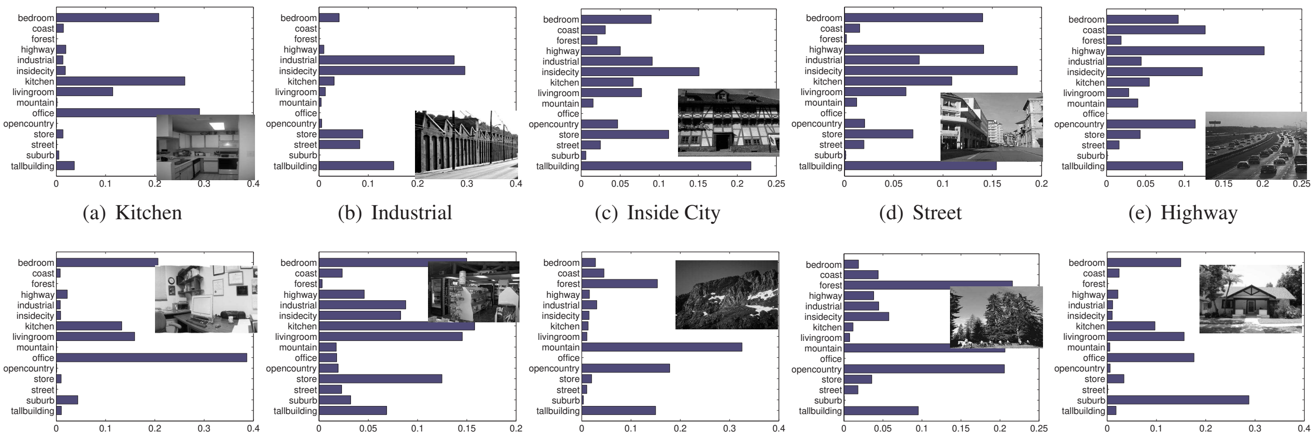
Formally, the count vector for the y^{th} image is an observation from a multinomial variable $\mathbf{T}$ of parameters $\boldsymbol{\pi}^{(y)} = (\pi_1^{(y)}, \ldots, \pi_L^{(y)})^T$

$$P_{\mathbf{T}|Y}(\mathcal{I}|y; \boldsymbol{\pi}^{(y)}) = \frac{n!}{\prod_{k=1}^{L} f_k!} \prod_{j=1}^{L} (\pi_j^{(y)})^{f_j}, \quad (1)$$

where $\pi_i^{(y)}$ is the probability that an image feature vector is drawn from the i^{th} theme. Note that this is a generic representation which can be implemented in many different ways. An implementation must simply specify a method to estimate the parameters $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_L)^T$ from the casually annotated training set. Each such vector $\boldsymbol{\pi}^{(y)}$ lies on an L-dimensional probability simplex, S_L, as shown in Fig. 1(left), which is a semantic feature space (i.e. its dimensions have a semantic interpretation). The scene classifier (e.g. SVM) then operates on this feature space. We next describe an implementation compatible with this generic framework.

3.2. Implementation Details

Although any semantic labeling system can be used to learn the semantic theme densities, we adopt the weakly supervised method of Carneiro et al. [4] as it is shown to achieve better performance than a number of other state-of-the-art methods available in the literature [7, 9]. The semantic theme density $P_{\mathbf{X}|T}(\mathbf{x}|t)$ is learned for each theme t_i from the set $\mathcal{D}_t$ of all training images annotated with the t^{th} caption in $\mathcal{L}$, using a *hierarchical estimation* procedure first proposed in [19], for image indexing. This procedure is itself composed of two steps as shown in Fig. 1(right). First,



(a) Kitchen (b) Industrial (c) Inside City (d) Street (e) Highway

(f) Office (g) Store (h) Mountain (i) Forest (j) Suburb

Figure 2. Some correctly classified images and their theme vectors, from 15-scene categories.

a Gaussian mixture is learned for each image in $\mathcal{D}_t$, producing a sequence of mixture densities $P_{\mathbf{X}|S,T}(\mathbf{x}|s,t)$, where S is a hidden variable that indicates the index of the image in $\mathcal{D}_t$. The second step is an extension of the EM algorithm, which clusters the Gaussian components of each image into a single mixture distribution, (see [4, 19] for details).

$$P_{\mathbf{X}|T}(\mathbf{x}|t;\Omega_t) = \sum_j \beta_t^j \mathcal{G}(\mathbf{x}, \nu_t^j, \mathbf{\Phi}_t^j) \tag{2}$$

It should be noted that the first step, learning of individual mixtures for all images in the training set, has complexity identical to that of learning a visterm codebook in the bag-of-visterms representation. The second step is extremely efficient, and has negligible complexity when compared with the first. This makes it much simpler than the unsupervised approaches previously used to learn latent spaces (LDA, pLSA, etc.), which frequently cannot be computed exactly and require variational approximations or Monte Carlo simulation.

Given an image $\mathcal{I} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$, the posterior theme probabilities

$$\pi_t = P_{T|\mathbf{X}}(t|\mathcal{I}) \tag{3}$$

are maximum a posteriori estimates of the parameters $\boldsymbol{\pi}_i$, and can be computed by combining (2) and Bayes rule, (assuming a uniform prior concept distribution $P_T(t)$), conditioned on the fact that all $\mathbf{x}_i$ are sampled independently.

4. Experimental evaluation

We now present an empirical evaluation of the proposed model for two publicly available datasets, comparing performance with [12, 3, 11, 10]. We also show that the theme vectors capture the semantic characteristics of most images in these datasets, and are efficiently able to learn theme co-occurrences. Finally a study of classification accuracy as a function of semantic space dimensions is presented.

4.1. Datasets

Scene classification results are presented on two public datasets: 1) 15-natural scene categories [10] and 2) Corel stock photos, used in [4] for image annotation comprising of 50 scene categories. The 15-scene categories contains 13 categories that were used by [11, 3]. The use of the 15-scene category dataset allow us to directly compare with the existing results on scene classification. In particular, we show a comparison of our results using low-dimensional representation with those of [12, 10, 11, 3]. The Corel dataset has 100 high resolution images per category. To the best of our knowledge, this is the database with maximum number of scene categories so far studied in the literature (viz. 50). Since the dimension of our semantic theme representation directly depends on the number of scene categories (see Sec. 3.1.2), this dataset enables the study of the effects of dimensionality as the number of categories grows.

4.2. Experimental Protocol

At the low level, images are represented as bags of 8×8 vectors of discrete cosine transform (DCT) coefficients sampled on a uniform grid. The Corel dataset consists of color images which are converted from RGB to YCrCb colorspace [3]. The 15-scene categories, consist of grayscale images hence no such conversion is required. Semantic theme densities are learned on a 36(out of 64) / 64(out of 192) dimensional subspace of the DCT coefficients for 15-scene categories and Corel dataset respectively, with each theme modeled as a mixture of 128 Gaussian components. The images at the semantic theme level are represented by 15 (50) dimensional theme vectors for 15-scene categories (Corel dataset). Later on, we also show that not all 50 themes are equally informative on Corel. 100 (90) images per scene are used to learn the theme density for 15-scene categories

[3] We also conducted experiments with the CIE lab colorspace and the results are almost similar.

	Office	Livingroom	Bedroom	Kitchen	Store	Industrial	TallBuilding	InsideCity	Street	Highway	Coast	Opencountry	Mountain	Forest	Suburb
Office	.96	.01	.02	.01	.00	.00	.01	.00	.00	.00	.00	.00	.00	.00	.00
Livingroom	.05	.55	.19	.05	.01	.00	.02	.03	.08	.00	.01	.01	.00	.00	.01
Bedroom	.09	.25	.36	.14	.03	.03	.01	.02	.04	.01	.00	.00	.02	.00	.00
Kitchen	.07	.07	.04	.66	.08	.05	.00	.02	.01	.00	.00	.00	.00	.00	.00
Store	.00	.02	.01	.07	.80	.08	.00	.00	.01	.00	.00	.00	.00	.00	.00
Industrial	.00	.00	.05	.04	.12	.69	.03	.01	.02	.00	.00	.01	.00	.01	.00
TallBuilding	.00	.02	.02	.00	.01	.04	.71	.05	.09	.02	.00	.00	.00	.01	.03
InsideCity	.00	.03	.01	.02	.01	.02	.06	.74	.07	.01	.00	.00	.00	.00	.01
Street	.00	.03	.03	.01	.02	.02	.10	.06	.66	.06	.00	.01	.00	.00	.02
Highway	.00	.01	.01	.00	.01	.02	.01	.02	.02	.78	.05	.06	.02	.00	.00
Coast	.00	.00	.00	.00	.00	.02	.00	.01	.00	.04	.77	.13	.02	.00	.00
Opencountry	.00	.00	.01	.00	.01	.00	.00	.00	.01	.04	.10	.66	.08	.08	.00
Mountain	.01	.00	.01	.00	.01	.01	.01	.00	.01	.01	.04	.10	.71	.07	.01
Forest	.00	.00	.00	.00	.04	.00	.01	.00	.01	.01	.00	.05	.06	.82	.01
Suburb	.00	.00	.00	.00	.00	.00	.02	.01	.00	.00	.00	.00	.00	.01	.96

Figure 3. Confusion Table for our method using 100 training image and rest as test examples from each category of 15-scene categories. The average performance is 72.2% $\pm$ 0.2

(Corel Dataset), and the rest of the images are used as the test set. All experiments on 15-scene categories are repeated 5 times with different randomly selected train and test images. For Corel dataset, we use the same training and test images as used in [4, 6]. A multi-class SVM using one-vs-all strategy with Gaussian kernel is used for classification, with the parameters obtained by 3-fold cross validation.

4.3. Results

We start by studying scene classification accuracy.

4.3.1 Scene classification

Fig. 2 shows some example images from the 15-scene categories, along with their semantic theme representation. All images shown are actually classified correctly by the classifier. Two interesting observations can be made: 1) semantic theme vectors *do capture* the different semantic meanings of the images, hence correlating well with human perception. For example, the theme vector shown for the scene from the category "Forest" in Fig. 2(i), has large weights for themes such as "*forest*", "*mountain*" and "*open-country*", which are suitable themes for the scene, and 2) in many examples (viz. Fig. 2(a)-(d),(g)), even though the semantic theme corresponding to the same semantic scene category does not have the highest probability, the scene is still classified correctly. For example in Fig. 2(d), in spite of the "*street*" theme having much lower probability than "*tall-building*", "*inside-city*", "*highway*", the image is classified as belonging to the "Street" category. This is a direct consequence of the classifier learning *associations* between themes, despite the casual nature of the annotations. Fig. 4 presents some of the misclassified images from the worst performing scene categories, along with the scene category they are classified into.

The confusion table for 15-scene categories is shown in Fig. 3. The average classification accuracy, over all categories is $72.2 \pm 0.2\%$. On Corel, the classification accuracy stands at 56.8%, the chance classification accuracy being 2%. Fig. 5 shows some of the images from various scene categories of Corel dataset.

→ Kitchen Bedroom (36%) | → Office Livingroom (55%) | → Inside City Street (66%) | → Store Kitchen (66%)

Figure 4. Some misclassified images from worst performing scenes in 15-scene categories. (→) implies the category image is classified into.

→ Trains (b) Trains | → Flowers (c) Peru | → Rural France (d) Africa | → Tropical Plants (e) Birds

Figure 5. Some images from the Corel dataset. (→) implies the category image is classified into.

4.3.2 Comparison with existing work

Table. 1 compares classification accuracy of the proposed method on 15-scene categories with existing results in the literature. It is evident that when compared to the MMI based dimensionality reduction of Liu et at. [12], which achieves a rate of 63.32% using a 20 dimensional space, the method performs substantially better, achieving a rate of 72.2% on an even lower dimensional space of 15 themes. Performance is equal to that of Lazebnik et al. [10][4], who represent images as the basic "bag-of-visterms" model, using 200 visterms. A similar comparison on the thirteen subcategories of the dataset used in [11, 3] is also presented in Table. 1.

4.3.3 Informative semantic themes

In all the experiments conducted above, scene categories served as a proxy for the intermediate themes. This is a practical approach to scene classification where the images are devoid of other annotations. However, it might seem that the extension of the current framework to very large-scale problems involving thousands of categories, will annul the benefits gained by the proposed representation, as the dimension of the semantic space would grow with the number of categories. The effects of varying the dimensions of the semantic space on the classification accuracy is

[4]Note that the best results on this dataset, are obtained by incorporating spatial information, and representing images as histograms at different spatial resolution, with Spatial Pyramid Matching [10]. The accuracy is 81.1%, with a 4200 dimensional feature space. However these extensions are beyond the scope of current discussion.

Table 1. Classification Result for 15 and 13 scene categories.

Method	Dataset	Dimensions	Accuracy
Our method	**15 Cat.**	**15**	**72.2 ± 0.2**
Liu et al. [12]	"	20	63.32
Liu et al. [12]	"	200	75.16
Lazebnik et al. [10]	"	200	72.2 ± 0.6
Our method	**13 Cat.**	**13**	**72.7 ± 0.3**
Bosch et al. [3]	"	25	73.4
Fei-Fei et al. [11]	"	40	65.2
Lazebnik et al. [10]	"	200	74.7

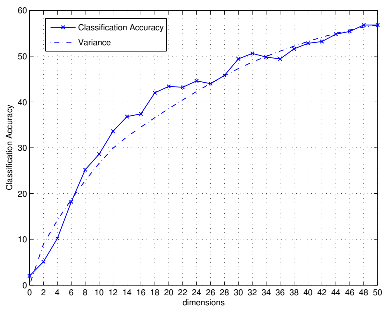


Figure 6. Classification performance as a function of the semantic space dimensions. Also shown, is the growth of the variance of the semantic themes, scaled appropriately.

studied, on Corel dataset. Semantic spaces of k dimensions were produced by ordering the semantic themes by the variance of their posterior probabilities, and selecting the k of largest variance (for k ranging from 2 to 50). Classification was performed on each of these resulting spaces and Fig. 6 presents the performance as a function of the dimension. It can be observed that not all of the 50 dimensions are equally informative, as moving from 40 to 50 dimensions increases performance by only 3.8% (a relative gain of 6.7%). This can be explained by the plot of variance of the posterior probabilities for the 50 themes (in the same figure). For very large scale problems, where most of the variance is expected to be captured by a subset of the features, the correlation of classification performance with the variance of the themes indicates that the number of informative themes would grow sub-linearly as the number of scene categories is increased. It is unclear that this type of behavior will hold for the flat bag-of-visterms representations. In the works previously presented in the literature, the codebook has *linear* size on the number of classes.

5. Discussion and Conclusion

The results presented above allow a number of conclusions. While low dimensional semantic representations are desirable for the reasons discussed in Section 1, previous approaches based on latent-space models have failed to match the performance of the flat bag-of-visterms model, which has high dimensionality. We have shown that this is indeed possible, with methods that have much lower complexity than the latent-space approaches previously proposed, but make better use of the available labeling information. We have also shown that the proposed method extracts meaningful semantic image descriptors, despite the casual nature of the training annotations, and is able to learn co-occurrences of semantic themes without explicit training for these. Finally a study of the effect of dimensionality on the classification performance was presented, and indicated that the dimensionality would grow sub-linearly with the number of scene categories. This could be a significant advantage over the flat bag-of-visterms models which, although successful for the limited datasets in current use, will likely not scale well when the class vocabulary increases.

References

[1] I. Biederman. Aspects and extension of a theory of human image understanding. *Computational processes in human vision: An interdisciplinary perspective*, New Jersey, 1988.

[2] D. Blei, A. Ng, and M. Jordan. Latent dirichlet allocation. *The Journal of Machine Learning Research*, 2003.

[3] A. Bosch, A. Zisserman, and X. Munoz. Scene classification via plsa. In *ECCV*, pages 517 – 30, Graz, Austria, 2006.

[4] G. Carneiro, A. Chan, P. Moreno, and N. Vasconcelos. Supervised learning of semantic classes for image annotation and retrieval. *IEEE PAMI*, March, 2007.

[5] G. Csurka, C. Bray, C. Dance, and L. Fan. Visual categorization with bags of keypoints. *Workshop on Statistical Learning in Computer Vision, ECCV*, pages 1–22, 2004.

[6] P. Duygulu, K. Barnard, N. Freitas, and D. Forsyth. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In *ECCV*, Denmark, 2002.

[7] S. Feng, R. Manmatha, and V. Lavrenko. Multiple bernoulli relevance models for image and video annotation. In *CVPR*, 2004.

[8] T. Hofmann. Probabilistic latent semantic indexing. *ACM SIGIR*, pages 50–57, 1999.

[9] V. Lavrenko, R. Manmatha, and J. Jeon. A model for learning the semantics of pictures. In *NIPS, Vancouver*, 2003.

[10] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. *CVPR*, 2005.

[11] F.-F. Li and P. Perona. A bayesian hierarchical model for learning natural scene categories. In *IEEE CVPR*, pages 524–531, 2005.

[12] J. Liu and M. Shah. Scene modeling using co-clustering. *ICCV*, 2007.

[13] E. Nowak, F. Jurie, and B. Triggs. Sampling strategies for bag-of-features image classification. *Proc. ECCV*, 4:490–503, 2006.

[14] 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.

[15] P. Quelhas, F. Monay, J.-M. Odobez, D. Gatica-Perez, T. Tuytelaars, and L. Van Gool. Modeling scenes with local descriptors and latent aspects. *ICCV*, Vol. 1:883 – 90, 2005.

[16] G. Salton and J. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill, New York, 1983.

[17] J. Sivic and A. Zisserman. Video google: a text retrieval approach to object matching in videos. *ICCV*, pages 1470–1477, 2003.

[18] M. Szummer and R. Picard. Indoor-outdoor image classification. In *IEEE Workshop on Content-based Access of Image and Video Databases*, 1998, Bombay, India.

[19] N. Vasconcelos. Image indexing with mixture hierarchies. In *Proc. IEEE CVPR*, Kawai, Hawaii, 2001.

[20] J. Vogel and B. Schiele. A semantic typicality measure for natural scene categorization. *DAGM04 Annual Pattern Recognition Symposium*.