Holistic Context Models for Visual Recognition

Nikhil Rasiwasia, *Student Member, IEEE,* Nuno Vasconcelos, *Senior Member, IEEE,*

***Abstract*— A novel framework to context modeling, based on the probability of co-occurrence of objects and scenes is proposed. The modeling is quite simple, and builds upon the availability of robust appearance classifiers. Images are represented by their posterior probabilities with respect to a set of contextual models, built upon the bag-of-features image representation, through two layers of probabilistic modeling. The first layer represents the image in a semantic space, where each dimension encodes an appearance-based posterior probability with respect to a concept. Due to the inherent ambiguity of classifying image patches, this representation suffers from a certain amount of contextual noise. The second layer enables robust inference in the presence of this noise, by modeling the distribution of each concept in the semantic space. A thorough and systematic experimental evaluation of the proposed context modeling is presented. It is shown that it captures the contextual "gist" of natural images. Scene classification experiments show that contextual classifiers outperform their appearance-based counterparts, irrespective of the precise choice and accuracy of the latter. The effectiveness of the proposed approach to context modeling is further demonstrated through a comparison to existing approaches on scene classification and image retrieval, on benchmark datasets. In all cases, the proposed approach achieves superior results.**

***Index Terms*— Computer Vision, Scene Classification, Context, Image Retrieval, Topic Models.**

I. INTRODUCTION

Visual recognition is a fundamental problem in computer vision. It subsumes the problems of scene classification [25], [26], [7], [37], [40], image annotation [9], [15], [24], [13], [5], image retrieval [12], [46], [39], [54], object recognition/localization [48], [44], [18], and object detection [56], [43], [16]. While the last decade has produced significant progress towards the solution of these problems, the basic strategy has remained constant: 1) identify a number of visual classes of interest, 2) design a set of appearance features (or some other visual representation, e.g., parts) that are discriminant for those classes, 3) postulate an architecture for their classification, and 4) rely on sophisticated statistical tools to learn optimal classifiers from training data. We refer to the resulting classifiers as *"appearance based"*. Main recent innovations produced better features, e.g. the ubiquitous SIFT descriptor [30], efficient classification architectures, namely the detector cascade of [56], methods for fast object matching [30], sophisticated discriminant classifiers, such as support vector machines (SVMs) with various kernels tunned for vision [19], [10], [7], [61], [8], and sophisticated statistical models [9], [26], [5], [48], [45], among others.

When compared to biological recognition strategies, strictly appearance-based methods have the limitation of not exploiting *contextual cues*. Psychophysics studies have shown that humans rarely guide recognition exclusively by the appearance of the concepts to recognize. Most frequently, appearance is complemented by the *analysis of contextual relationships* with other visual concepts in the field of view [4]. In general, the detection of a concept of interest (e.g. buildings) is facilitated by the presence, in the scene, of other concepts (e.g. street, city) which *may not* themselves be of interest. Psychophysical studies have shown that context can depend on multiple clues. For example, object recognition is known to be affected by properties such as support (objects do not float in the air), interposition (objects occupy different volumes), probability (objects appear in different scenes with different probabilities), position (objects appear in typical locations), and size (objects have typical relative sizes) [4].

In this work, we investigate an approach to context modeling based on the probability of co-occurrence of objects and scenes. This modeling is quite simple, and builds upon the availability of robust appearance classifiers. A vocabulary of visual concepts is defined, and statistical models learned for all concepts, with existing appearance modeling techniques [9], [25], [26]. These techniques are typically based on the *bag-of-features* (BoF) representation, where images are represented as collections of spatially localized features. The outputs of the appearance classifiers are then interpreted as the dimensions of a *semantic space*, where each axis represents a visual concept [39], [57], [47], [32]. This is illustrated in Figure 1, where an image is represented by the vector of its posterior probabilities under each of the appearance models. This vector is denoted as a *semantic multinomial* (SMN) distribution [39]. An example SMN for a natural image is shown in Figure 3 (bottom).

This semantic representation inherits many of the benefits of bag-of-features. Most notably, it is strongly invariant to scene configurations, an essential attribute for robust scene classification and object recognition, and has low complexity, a property that enables large training sets and good generalization. Its main advantage over bag-of-features is a higher level of *abstraction*. While the appearance features are edges, edge orientations, or frequency bases, those of the semantic representation are concept probabilities. We have previously shown that this can lead to substantially better generalization, by comparing the performance of nearest-neighbors classification with the two representations, in an image retrieval context [39]. However, the semantic representation also has some limitations that can be traced back to bag-of-features. Most notable among these is a certain amount of *contextual noise*, i.e., noise in the probabilities that compose the SMN. This is usually not due to poor statistical estimation, but due to the intrinsic *ambiguity* of the underlying bag-of-features representation. Since appearance based features have small spatial support, it is frequently difficult to assign them to a single visual concept. Hence, the SMN extracted from an image usually assigns some probability to concepts unrelated to it (e.g. the concepts "bedroom" and "kitchen" for the "street" image of Figure 3).

While the SMN representation captures co-occurrences of the semantic concepts present in an image, not all these correspond to *true* contextual relationships. In fact, we argue that many (e.g.

• *N. Rasiwasia and N. Vasconcelos are with the Department of Electrical and Computer Engineering, University of California, San Diego, 9500 Gilman Dr., La Jolla, CA 92093-0407. E-mail: {nikux,nvasconcelos}@ucsd.edu*

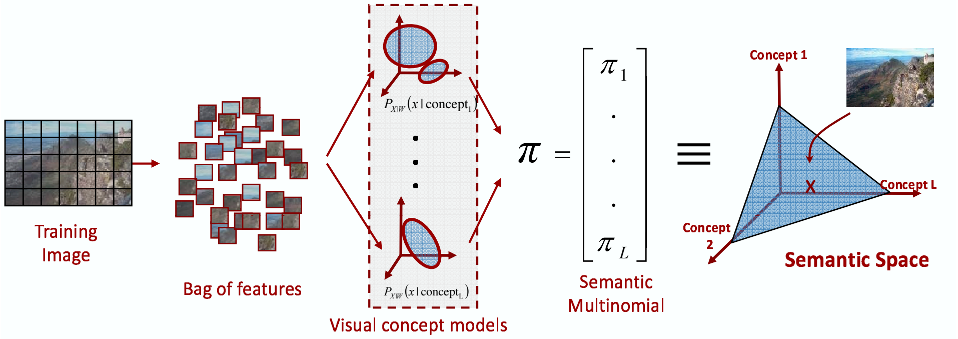

Training Image
Bag of features
$P_{X|W}(x\,|\,\text{concept}_1)$
$P_{X|W}(x\,|\,\text{concept}_L)$
Visual concept models
$\pi = \begin{bmatrix} \pi_1 \\ . \\ . \\ . \\ \pi_L \end{bmatrix} \equiv$
Semantic Multinomial
Concept 1
Concept L
Concept 2
X
Semantic Space

Fig. 1. Image representation in semantic space $\mathcal{S}$, with a semantic multinomial (SMN) distribution. The SMN is a vector of posterior concept probabilities which encodes the co-occurrence of various concepts in the image, based on visual appearance.

"bedroom" and "kitchen" in Figure 3) are *accidental*, i.e., casual coincidences due to the ambiguity of the underlying appearance representation (image patches that could belong to either a bed or a kitchen counter). Rather than attempting to eliminate contextual noise by further processing of appearance features, we propose a procedure for *robust* inference of contextual relationships *in the presence of accidental co-occurrences*. The idea is to keep the robustness of the appearance representation, but perform the classification at a higher level of *abstraction*, where ambiguity can be more easily detected.

This is achieved by introducing a second level of representation, that operates in the space of semantic features. The intuition is that, in this space, accidental co-occurrences are events of much smaller probability than true contextual co-occurrences: while "street" co-occurs with "buildings" in most images, it accidentally co-occurs with "bedroom" or "kitchen" in a much smaller set. True contextual relationships can thus be found by identifying peaks of probability in semantic space. Each visual concept is modeled by the distribution of the posterior probabilities extracted from all its training images. This *distribution of distributions* is referred as the *contextual model* for the concept. For large enough and diverse enough training sets, these models are dominated by the probabilities of true contextual relationships. Minimum probability of error (MPE) contextual classification can thus be implemented by simple application of Bayes' rule. This suggests representing images as vectors of posterior probabilities under the contextual concept models, which we denote by *contextual multinomials* (CMN). These are shown much less noisier than the SMNs learned at the appearance level.

An implementation of contextual modeling is proposed, where concepts are modeled as mixtures of Gaussian distribution on appearance space, and mixtures of Dirichlet distributions on semantic space. It is shown that 1) the contextual representation outperforms the appearance based representation, and 2) this holds irrespectively of the choice and accuracy of the underlying appearance models. An extensive experimental evaluation, involving the problems of scene classification and image retrieval shows that, despite its simplicity, the proposed approach is superior to various contextual modeling procedures in the literature.

The paper is organized as follows. Section II briefly reviews the literature on context modeling. Our previous work on appearance classification and the design of semantic spaces is reviewed on Section III. Section IV then discusses the limitations of appearance classifiers and introduces contextual models. Some practical issues in the design of the latter are discussed in Section V. Section VI relates the architecture now proposed to the literature on topic models. An extensive experimental evaluation of contextual modeling is then presented in Sections VII, VIII, and IX. Finally, Section X presents some conclusions. A preliminary version of this work appeared in [41].

II. Context Modeling

Recent efforts towards context based recognition can be broadly grouped in two classes. The first, an *object-centric* approach, consists of methods that model contextual relationships between sub-image entities, such as objects. Examples range from simply accounting for the co-occurrence of different objects in a scene [38], [17], to explicit learning of the spatial relationships between objects [18], [60], or an object and its neighboring image regions [20]. Methods in the second class adopt a *scene-centric* representation, whereby context models are learned from entire images, generating a holistic description of the scene or its "gist" [34], [57], [26], [35], [25]. Various recent works have shown that semantic descriptions of natural images can be obtained with these representations, without explicit image segmentation [34]. This is consistent with evidence from the psychology [33] and cognitive neuroscience [1] literatures.

The scene-centric representation has itself been explored in two ways. One approach is to equate context to a vector of statistics of low-level visual measurements taken over the entire image. For example, [34] models scenes according to the differential regularities of their second order statistics. A second approach is to rely on the bag-of-features representation. Here, low-level features are computed locally and aggregated across the image, to form a holistic context model [57], [26], [42]. Although these methods usually ignore spatial information, some extensions have been proposed to weakly encode the latter. These consist of dividing the image into a coarse grid of spatial regions, and modeling context within each [34], [25].

The proposed context modelling combines aspects of both the object-centric and scene-centric strategies. Like the object-centric methods, we exploit relationships between co-occurring semantic concepts in natural scenes to derive contextual information. This is, however, accomplished without demarcating individual concepts or regions in the image. Instead, all conceptual relations are learned through global scene representations. Moreover, these relationships are learned in a purely data-driven fashion, i.e. no external guidance about the statistics of high-level contextual relationships is required, and the representation consists of full probability distributions, not just statistics. The proposed representation can be thought as modeling the "gist" of the scene by the co-occurrences of semantic visual concepts that it contains.

The representation closest to that now proposed is probably the family of latent topic models, recently popular in vision [26], [37], [7]. These models were originally proposed in the text literature, to address the ambiguity of bag-of-words. It was realized that word histograms cannot account for polysemy (the same word may represent different meanings) and synonymy (different words may represent same meaning) [6], [21]. This led to the introduction of intermediate latent representations, commonly known as "themes" or "topics". Borrowing from the text literature, several authors applied the idea of latent spaces to visual recognition [5], [2], [45], [48], [26], [37], [7]. The rational is that images which share frequently co-occurring features have a similar representation in the latent space. Although successful for text, the benefits of topic discovery have not been conclusively established for visual recognition. In fact, a drop in classification performance is often experienced when unsupervised latent representations are introduced [28], [37], [25]. This issue is discussed in detail in Section VI, where we argue that unsupervised topic discovery is not a good idea for recognition. We show that the architecture now proposed can be interpreted as a modified topic model, where the topics are pre-specified and learned in a weakly supervised manner. This is shown to increase the recognition performance.

The use of appearance based classifier outputs as feature vectors has also been proposed in [40], [58], [51]. In these works a classifier is first learned for a given keyword vocabulary — [58], [51] learn discriminative classifiers from `flickr/bing` images, [40] learns a generative model using a labeled image set — and the outputs of these classifiers are then used as feature vectors for a second layer of classification. In these works, classifier outputs are simply used as an alternative low dimensional image representation, without any analysis of their ability to model context. We discuss the limitations of using appearance models for context modeling and introduce "contextual models" that address these limitations. We also present extensive experimental evidence supporting the benefits of these higher level models, and show that they achieve higher classification accuracies on benchmark datasets.

III. Appearance-based Models and Semantic Multinomials

We start by briefly reviewing appearance-based modeling and the design of semantic spaces.

A. Notations

Images are observations from a random variable $\mathbf{X}$, defined on some feature space $\mathcal{X}$ of visual measurements. For example, $\mathcal{X}$ could be the space of discrete cosine transform (DCT), or SIFT descriptors. Each image is represented as a bag of N *feature vectors* $\mathcal{I} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}, \mathbf{x}_i \in \mathcal{X}$ assumed to be sampled independently. A collection of images is called an "image dataset", $\mathcal{D} = \{\mathcal{I}_1, \ldots, \mathcal{I}_D\}$.

Each image is labeled with a label vector, $\mathbf{c}_d$ according to a vocabulary of semantic concepts $\mathcal{L} = \{w_1, \ldots, w_L\}$ making $\mathcal{D} = \{(\mathcal{I}_1, \mathbf{c}_1), \ldots, (\mathcal{I}_D, \mathbf{c}_D)\}$. Note that $\mathbf{c}_d$ is a binary L-dimensional vector such that $c_{d,i} = 1$ if the d^{th} image was annotated with the i^{th} keyword in $\mathcal{L}$. The dataset is said to be weakly labeled if absence of a keyword from caption $\mathbf{c}_d$ does not necessarily mean that the associated concept is not present in $\mathcal{I}_d$. For example, an image containing "sky" may not be explicitly labeled with that keyword. This is usually the case in practical scenarios, since each image is likely to be annotated with a small caption that only identifies the semantics deemed as most relevant to the labeler. In fact, for certain recognition tasks, such as scene classification or image retrieval, an image is usually annotated with just one concept. We assume weak labeling throughout this work.

B. Appearance-based Classification

Visual concepts are drawn from a random variable W, which takes values in $\{1, \ldots, L\}$. Each concept w induces a probability density on $\mathcal{X}$, which is approximated by a model $P_{\mathbf{X}|W}(\mathbf{x}|w)$. This is denoted as the appearance model for concept w, and describes how observations are drawn from this concept, in the low-level visual feature space $\mathcal{X}$. $P_{\mathbf{X}|W}(\mathbf{x}|w)$ is learned from the set $\mathcal{D}_w$ of all training images whose caption includes the w^{th} label.

Many appearance recognition systems have been proposed in the literature, using different appearance models. A simple generative model for appearance is shown in Figure 2. A concept w is first sampled, and N feature vectors are then generated from the class-conditional distribution $P_{\mathbf{X}|W}(\mathbf{x}|w)$. This model performs well in weakly supervision concept detection problems [9]. Given an unseen image $\mathcal{I}$, MPE detection is achieved with the Bayes decision rule

$$P_{W|\mathbf{X}}(i|\mathcal{I}) = \frac{P_{\mathbf{X}|W}(\mathcal{I}|i)P_W(i)}{P_{\mathbf{X}}(\mathcal{I})} \tag{1}$$

$$= \frac{\prod_{n=1}^{N} P_{\mathbf{X}|W}(\mathbf{x}_n|i)P_W(i)}{\prod_{n=1}^{N} P_{\mathbf{X}}(\mathbf{x}_n)} \tag{2}$$

where $P_{W|\mathbf{X}}(i|\mathcal{I})$ is the probability of presence of the i^{th} concept in the image, given the observed set of feature vectors $\mathcal{I}$. We assume a uniform prior concept distribution $P_W(w)$, although any other suitable prior could be used. This leads to

$$P_{W|\mathbf{X}}(i|\mathcal{I}) \propto \frac{\prod_{n=1}^{N} P_{\mathbf{X}|W}(\mathbf{x}_n|i)}{\prod_{n=1}^{N} P_{\mathbf{X}}(\mathbf{x}_n)} \tag{3}$$

To model the appearance distribution, we rely on Gaussian mixture models (GMM). These are popular models for the distribution of visual features [9], [20], [49], [5] and have the form

$$P_{\mathbf{X}|W}(\mathbf{x}|w; \Omega^w) = \sum_k \rho_k^w \mathcal{G}(\mathbf{x}, \mu_k^w, \Sigma_k^w), \tag{4}$$

where $\Omega^w = \{\rho_k^w, \mu_k^w, \Sigma_k^w\}$, ρ_k^w is a probability mass function such that $\sum_k \rho_k^w = 1$, and $\mathcal{G}(\mathbf{x}, \mu, \Sigma)$ a Gaussian density of mean μ and covariance Σ. The parameters Ω^w are learned with a *hierarchical estimation* procedure first proposed in [55], for image indexing (see [9], [55] for details).

C. Designing a Semantic Space

While the Bayes decision rule for concept detection only requires the largest posterior concept probability for a given image, the vector of posterior probabilities $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_L)^T$, where $\pi_w = P_{W|\mathbf{X}}(w|\mathcal{I})$ provides a rich description of the image semantics. We refer to this vector as a *semantic multinomial* (SMN) distribution, which lies on a probability simplex $\mathcal{S}$, referred to as the *semantic space* [39]. As shown in Figure 1, this alternative

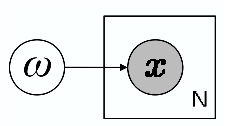


Fig. 2. The generative model underlying image formation at the visual level. w represents a sample from a vocabulary of semantic concepts, and an image $\mathcal{I}$ is composed of N patches, $\mathbf{x}_n$, sampled independently from $P_{\mathbf{X}|W}(\mathbf{x}|w)$. Note that, throughout this work, we adopt the standard plate notation of [6] to represent graphical models.

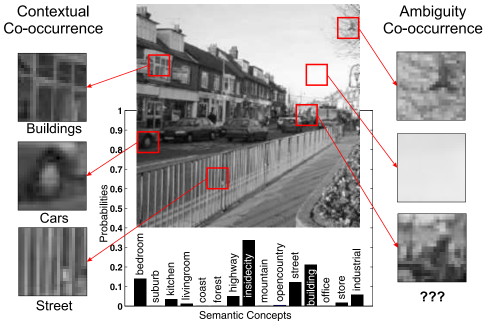


Fig. 3. An image from the "street" class of the N15 dataset (See Section. VII-A) along with its SMN. Also highlighted are the two notions of *co-occurrence*. *Ambiguity co-occurrences* on the right: image patches compatible with multiple unrelated classes. *Contextual co-occurrences* on the left: patches of multiple other classes related to "street".

image representation establishes a mapping from images in $\mathcal{X}$ to SMNs $\boldsymbol{\pi}$ in $\mathcal{S}$. This can be seen as an abstract projection of the image onto a space where each concept probability π_w, $w = 1, \ldots, L$ is a feature. Unlike $\mathcal{X}$, these features have explicit semantics. Thus, while inheriting many of the benefits of bag-of-features, such as invariance to scene configuration and low complexity, the semantic representation has the advantage of a higher level of *abstraction*. While appearance features are edges, edge orientations, or frequency bases, those of the semantic representation are concept probabilities. This representation has been shown successful for image retrieval, where images are matched using a nearest neighbor operation on the semantic space [39], [47], [32]. Nevertheless, it is not free of limitations.

D. Limitations of Semantic Representations

One major source of difficulties is that semantic models built upon the bag-of-features representation of appearance inherit the ambiguities of the latter. There are two main types of ambiguity. The first is that contextually unrelated concepts (for example smoke and clouds) can have similar appearance representation under bag-of-features. The second is that the resulting semantic descriptors can account for contextual frequencies of co-occurrence, but not true contextual dependencies. These two problems are illustrated in Figure 3. First, image patches frequently have ambiguous interpretation. When considered in isolation, they can be compatible with many concepts. For example it is unclear that even a human could confidently assign the patches shown on the right of Figure 3 to the "street" concept, with which the image is labeled. Second, appearance-based models lack information about the interdependence of the semantics of the patches which compose the images in a class. For example, the fact that, as shown on the left, images of street scenes typically contain patches of street, car wheels, and building texture.

We refer to these two observations as *co-occurrences*. In the first case, a patch can accidentally co-occur with multiple concepts (the equivalent to *polysemy* in text analysis). In the second, patches from multiple concepts typically co-occur in scenes of a given class (the equivalent to *synonymy* for text). While only the co-occurrences of the second type are indicative of *true* contextual relationships, SMNs learned from appearance models capture *both* types of co-occurrences. This is again illustrated by the example of Figure 3. On one hand, the displayed SMN reflects the *ambiguity* that sometimes exists between patches of "street scenes" and "bedrooms", "kitchens" or "living rooms". These are all man-made structures which, for example, contain elongated edges dues to buildings, beds, furniture, etc. Note that all classes that typically do not have such structures (e.g. natural scenes such as "mountain", "forest", "coast", or "open country") receive close to zero probability. On the other, the SMN reflects the likely co-occurrence, in "street scenes", of patches of "inside city", "street", "buildings", and "highway". In summary, while SMN probabilities can be interpreted as semantic features, which account for co-occurrences due to both ambiguity and context, they are not purely *contextual features*.

IV. Semantics-based Models and Context Multinomials

One possibility to deal with the ambiguity of the semantic representation is to explicitly model contextual dependencies. This can be done by introducing *constraints* on the appearance representation, by modeling constellations of parts [16], [14] or object relationships [50], [18]. However, the introduction of such constraints increases complexity, and reduces the invariance of the representation, sacrificing generalization. A more robust alternative is to keep bag-of-features, but represent images at a higher level of *abstraction*, where ambiguity can be more easily detected. This is the strategy pursued in this work, where we exploit the fact that the two types of SMN co-occurrences have different *stability*, to extract *more reliable* contextual features.

A. From Semantics to Context

The basic idea is that, while images from the same concept are expected to exhibit similar contextual co-occurrences, this is not likely for ambiguity co-occurrences. Although the "street scene" of Figure 3 contains some patches that could also be attributed to the "bedroom" concept, it is unlikely that this will hold for most images of street scenes. By definition, ambiguity co-occurrences are *accidental*, otherwise they would reflect common semantics of the two concepts, and would be contextual co-occurrences. Thus, while impossible to detect from a single image, stable contextual co-occurrences should be detectable by joint inspection of *all* SMNs derived from the images of a concept.

This is accomplished by extending concept modeling by one further layer of semantic representation. As illustrated in Figure 4, each concept w is modeled by the probability distribution of the SMNs derived from all training images in its training set, $\mathcal{D}_w$. We refer to this SMN distribution as the *contextual model* for w. If $\mathcal{D}_w$ is large and diverse, this model is dominated by the stable properties of the features drawn from concept w. In this case, the features are SMNs and their stable properties are the true contextual relationships of w. Hence, concept models assign high probability to regions of the semantic space occupied

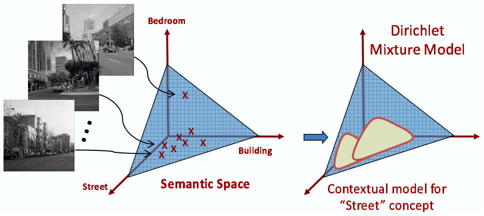


Fig. 4. Learning the contextual model for the "street" concept, (5), on semantic space $\mathcal{S}$, from the set of all training images annotated with "street".

by contextual co-occurrences, and small probability to those of ambiguity co-occurrences.

For example, since streets typically co-occur with buildings, the contextual model for "street" assigns high probability to SMNs that include both concepts. On the other hand, because "street" only co-occurs accidentally with "bedroom", SMNs including this concept receive low-probability. Hence, representing images by their posterior distribution under contextual models emphasizes contextual co-occurrences, while suppressing accidental coincidences due to ambiguity. As a parallel to the nomenclature of the previous section, we refer to the posterior probabilities at this higher level of abstraction as *contextual features*, the probability vector associated with each image as a *contextual multinomial* distribution, and the space of such vectors as the *contextual space*.

B. Contextual Concept Models

Contextual concept models are learned in the semantic space $\mathcal{S}$. Under the most general formulation, concepts are drawn from a random variable Y defined on the index set $y \in \{1, \ldots, K\}$ of a concept vocabulary $\mathcal{K}$. In this work, we assume that this vocabulary is the concept vocabulary $\mathcal{L}$ used in visual space $\mathcal{X}$, i.e. $\mathcal{K} = \mathcal{L}$. Note that this assumption implies that if $\mathcal{L}$ is composed of scenes (objects), then the contextual models account for relationships between scenes (objects). A trivial extension would be to make concepts on semantic space $\mathcal{S}$ different from those on visual space $\mathcal{X}$, promoting a concept hierarchy. For example, Y could be defined on the vocabulary of scenes, $\mathcal{K} = \{`desert', `beach', `forest'\}$ and $W$ on objects, $\mathcal{L} = \{`sand', `water', `sky', `trees'\}$. In this way, scenes in $\mathcal{K}$ would be naturally composed of objects in $\mathcal{L}$, enabling the contextual models to account for relationships between scenes and objects. This would, however, require training images (weakly) labeled with respect to both $\mathcal{L}$ and $\mathcal{K}$. We do not pursue such hierarchical concept taxonomies in what follows.

Since $\mathcal{S}$ is itself a probability simplex, one natural model for a concept y in $\mathcal{S}$ is the mixture of Dirichlet distributions

$$P_{\mathbf{\Pi}|Y}(\boldsymbol{\pi}|y;\Lambda^y) = \sum_k \beta_k^y \mathcal{D}ir(\boldsymbol{\pi};\boldsymbol{\alpha}_k^y). \tag{5}$$

This model has parameters $\Lambda^y = \{\beta_k^y, \boldsymbol{\alpha}_k^y\}$, where β_k is a probability mass function ($\sum_k \beta_k^y = 1$). $\mathcal{D}ir(\boldsymbol{\pi};\boldsymbol{\alpha})$ a Dirichlet distribution of parameter $\boldsymbol{\alpha} = \{\alpha_1, \ldots, \alpha_L\}$,

$$\mathcal{D}ir(\boldsymbol{\pi};\boldsymbol{\alpha}) = \frac{\Gamma(\sum_{i=1}^L \alpha_i)}{\prod_{i=1}^L \Gamma(\alpha_i)} \prod_{i=1}^L (\pi_i)^{\alpha_i - 1} \tag{6}$$

and $\Gamma(.)$ the Gamma function. As illustrated in Figure 4, the parameters Λ^y are learned from the SMNs $\boldsymbol{\pi}$ of all images in $\mathcal{D}_y$, i.e. the images annotated with the y^{th} concept in $\mathcal{L}$. Learning is implemented by maximum likelihood estimation, using the generalized expectation-maximization (GEM) algorithm discussed in Appendix I.

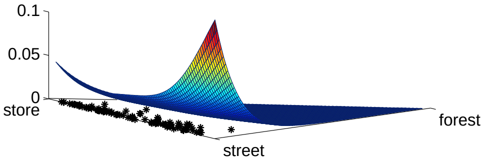


Fig. 5. 3-component Dirichlet mixture learned for the concept "street". Also shown, as "*", are the SMNs associated with each image. The Dirichlet mixture assigns high probability to the concepts "street" and "store".

Figure 5 shows an example of a 3-component Dirichlet mixture learned for the semantic concept "street", on a three-concept semantic space. This model is estimated from 100 images (shown as data points on the figure). Note that, although some of the image SMNs exhibit ambiguity co-occurrences with the "forest" concept, the Dirichlet mixture is strongly dominated by the true contextual co-occurrences between the concepts "street" and "store". This is an illustration of the ability of the model to lock onto the true contextual relationships.

C. Contextual Space

The contextual models $P_{\Pi|Y}(\boldsymbol{\pi}|y)$ play, in semantic space $\mathcal{S}$, a similar role to that of the appearance models $P_{\mathbf{X}|W}(\mathbf{x}|w)$ in visual space $\mathcal{X}$. It follows that MPE concept detection, on a test image $\mathcal{I}$ of SMN $\boldsymbol{\pi} = \{\pi_1, \ldots, \pi_L\}$, can be implemented with a Bayes decision rule based on the posterior concept probabilities

$$P_{Y|\Pi}(y|\boldsymbol{\pi}) = \frac{P_{\Pi|Y}(\boldsymbol{\pi}|y)P_Y(y)}{P_\Pi(\boldsymbol{\pi})}. \tag{7}$$

This is the semantic space equivalent of (2) and, once again, we assume a uniform concept prior $P_Y(y)$.

As in Section III-C, it is also possible to design a new semantic space, by retaining all posterior contextual concept probabilities $\theta_y = P_{Y|\Pi}(y|\boldsymbol{\pi})$. We denote the vector $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_L)^T$ as the *contextual multinomial* (CMN) distribution of image $\mathcal{I}$. As illustrated in Figure 6, CMN vectors lie on a new probability simplex $\mathcal{C}$, here referred to as the *contextual space*. In this way, the contextual representation establishes a mapping from images in $\mathcal{X}$ to CMNs $\boldsymbol{\theta}$ in $\mathcal{C}$. In Section VII we show that CMNs are much more reliable contextual descriptors than SMNs.

V. Learning Contextual Models

For now, we discuss a number of issues in the implementation of the architecture introduced above.

A. Computing the Semantic Multinomials

It should be noted that this architecture is generic, in the sense that any appearance recognition system that produces a vector of posterior probabilities $\boldsymbol{\pi}$, can be used to learn the proposed contextual models. In fact, these probabilities can even be produced by systems that do not model appearance explicitly, e.g. discriminant classifiers. This is achieved by converting classifier scores to a posterior probability distribution, using probability calibration techniques. For example, the distance from the decision hyperplane learned by an SVM can be converted to a

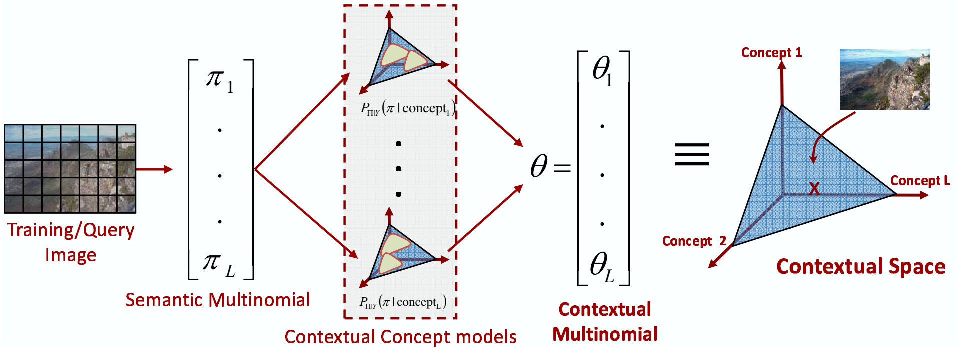


Fig. 6. The Contextual multinomial (CMN) of an image as the vector of co-occurrence probabilities of contextually related concepts.

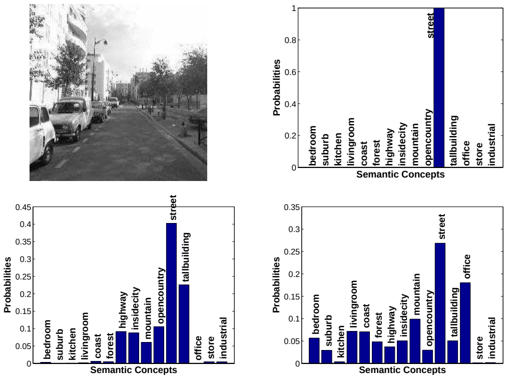


Fig. 7. SMN for the image shown on the top left computed using (top-right) (3), (bottom-left) (10) and (bottom-right) (12).

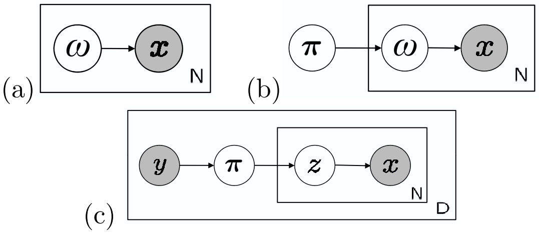


Fig. 8. Alternative generative models for image formation at the appearance level. (a) A concept is sampled per appearance feature vector rather than per image, from $P_{\mathbf{X}|W}(\mathbf{x}|w)$. (b) Explicit modeling of the contextual variable Π from which a single SMN is drawn per image. (c) Graphical model of LDA with an additional class variable [26].

posterior probability using a sigmoidal transform [36]. In practice, however, care must be taken to guarantee that the appearance classifiers are not too strong. If they make very hard decisions, e.g. assign images to a single class, little is left for the context models to work with. In this case, contextual processing fails to add any improvement to the appearance classification. This is a manifestation of the data processing theorem [31] which advises to postpone hard decisions until the very last stages of processing.

In the implementation above, it is natural to use the posterior probabilities of (3) as the SMN of image $\mathcal{I}$. However, N tends to be large, and there is usually very strong evidence in favor of one concept, not always that of greatest perceptual significance. For example, if the image has a large region of "sky", the existence of many sky patches makes the posterior probability of the "sky" concept close to one. This is illustrated in Figure 7 (top-right) where the SMN assigns all probability to a single concept. Table I shows that this happens frequently: the average entropy of the SMNs computed on the N15 Dataset (to be introduced later) is very close to 0. Note that this is the property that enables the learning of the appearance based models from the weakly supervised datasets: when all images containing "sky" are grouped, the overall feature distribution is very close to that of the "sky" concept, despite the fact that the training set contains all sorts of spurious image patches from other concepts. This is an example of the multiple instance learning paradigm [53], where an image, consisting of some patches from the concept being modeled and some spurious patches from other concepts, serves as the positive bag. Although this dominance of the strongest concept is critical for learning, the data processing theorem advises against it during inference. Or, in other words, while multiple instance learning is required, multiple instance inference is undesirable. In particular, modeling images as bags-of-features from a *single concept*, as in Figure 2, does not lend to contextual inference.

One alternative is to perform inference with the much looser model of Figure 8(a), where a concept is sampled *per appearance feature vector*, rather than per image. Note that, because labeling information is not available per vector, the models $P_{\mathbf{X}|W}(\mathbf{x}|w)$ are still learned as before. The only difference is the inference procedure. In this case, SMNs are available per image patch denoted as patch-SMN, $\boldsymbol{\pi}^n = P_{W|X}(w_n|x_n), n \in \{1, \dots, N\}$. Determining an SMN, denoted the *Image-SMN*, for the entire image requires computing a representative for this set of patch-SMNs. One possibility is the multinomial of minimum average Kullback-Leibler divergence with all patch-SMNs

$$\boldsymbol{\pi}^* = \arg\min_{\boldsymbol{\pi}} \frac{1}{N}\sum_{n=1}^{N} KL(\boldsymbol{\pi}||\boldsymbol{\pi}^n) \quad \text{s.t} \sum_{i=1}^{L} \pi_i = 1. \tag{8}$$

As shown in Appendix II, this is the representative

$$\pi_i^* = \frac{\exp \frac{1}{N}\sum_n \log \pi_i^n}{\sum_i \exp \frac{1}{N}\sum_n \log \pi_i^n}, \tag{9}$$

which reduces to

$$\pi_i^* = \frac{\exp\left\{\frac{1}{n}\sum_n \log P_{X|W}(x_n|i)\right\}}{\sum_j \exp\left\{\frac{1}{n}\sum_n \log P_{X|W}(x_n|j)\right\}} \tag{10}$$

for a uniform prior. This is in contrast to the posterior estimate of (3). Note that while (3) computes a product of likelihoods, (10) computes their geometric mean.

TABLE I
SMN ENTROPY.

Model	Entropy
Figure 2, Eq (3)	0.003 ± 0.044
Figure 8(a), Eq (10)	2.530 ± 0.435
Figure 8(b), Eq (12)	2.546 ± 0.593

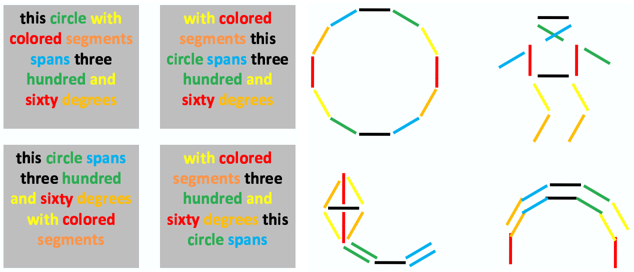

this circle with colored segments spans three hundred and sixty degrees
with colored segments this circle spans three hundred and sixty degrees
this circle spans three hundred and sixty degrees with colored segments
with colored segments three hundred and sixty degrees this circle spans

Fig. 9. left) Four groups of words with equal word histograms. right) Four groups of edge segments with the equal edge segment histograms. Note that each group can be derived from the others by a displacement of words or edge segments. (This figure is best viewed in color)

A second possibility is to adopt the generative model of Figure 8(b). This explicitly accounts for the contextual variable $\mathbf{\Pi}$, from which a single SMN is drawn per image. A concept is then drawn per image patch. In this case, the Image-SMN is

$$\boldsymbol{\pi}^* = \arg\max_{\boldsymbol{\pi}} P_{\Pi|X}(\boldsymbol{\pi}|\mathcal{I}). \tag{11}$$

However, this optimization is intractable, and only approximate inference is possible. A number of approximations can be used, including Laplace or variational approximations, sampling, etc. In Appendix III we show that, for a variational approximation,

$$\pi_i^* = \frac{\gamma_i - 1}{\sum_j \gamma_j - L} \tag{12}$$

where, γ_i is computed with the following iteration,

$$\gamma_i^* = \sum_n \phi_{ni} + \alpha_i \tag{13}$$

$$\phi_{ni}^* \propto P_{X|W}(x_n|w_n = i)\ e^{\psi(\gamma_i) - \psi(\sum_j \gamma_j)}. \tag{14}$$

Here, α_i is the parameter of the prior $P_\Pi(\pi)$ which, for compatibility with the assumption of uniform class priors, we set to 1, $\psi(\cdot)$ the Digamma function, and γ_i, ϕ_{ni} the parameters of the variational distributions. Figure 7 shows that the SMNs obtained with (10) and (12) are rich in contextual information. Table I shows that the two models lead to approximately the same average SMN entropy on N15, which is much higher than that of (3).

B. Data Augmentation

When, as above, an SMN is computed per image, the number of training images upper bounds the cardinality of the training set for contextual models. Since there is usually a limited number of labelled images per concept, this can lead to over fitting. For example, the 100 images available per concept on N15 are sufficient to learn appearance models (each image contains thousands of patches), but 100 SMNs do not suffice to learn Dirichlet mixtures in a 15 dimensional space. One possibility is to use the patch-SMNs, $\boldsymbol{\pi}^{(n)}$, which are abundant. These, however, tend to be too noisy, due to the ambiguities discussed above. To overcome this problem we resort to a middle ground between patch-SMNs and image-SMNs: multiple SMNs are estimated per image, from random patch subsets. More precisely, a set of patches is first selected, randomly, from the image. An SMN is then estimated from this set, as would be done if the image consisted of these patches alone. The process is repeated with different patch subsets, generating a number of SMNs per image. By controlling the number of random sets, it is possible to control the cardinality of the training set for each contextual model. The use of random patch subsets simultaneously alleviates the problems of data scarcity (many subsets can be drawn per image), and estimation noise (each SMN pools information from multiple patches). Moreover, similar to the learning of appearance models, learning contextual models with data augmentation also relies on the multiple instance learning paradigm where each image, being a collection of SMNs, serves as the positive bag, with some SMNs depicting true contextual co-occurrences and some others ambiguity co-occurrences. In Section VIII-A, we show that this data augmentation strategy leads to significant improvements in classification accuracy.

VI. CONNECTIONS TO TOPIC MODELS

The architecture proposed above has several properties in common with the family of *topic models,* [6], [21].

A. Topic Models

Like the representation now proposed, topic models have two layers. Appearance features are used to compute topic probabilities (that correspond to the proposed SMNs), which are hierarchically propagated to a more abstract layer that computes class probabilities (correspondent to the proposed CMNs). While details vary, the models are usually variations of latent Dirichlet allocation (LDA) [6], or probabilistic latent semantic analysis (pLSA) [21]. [5], [2] present an extension of LDA for image annotation. Two other variations are proposed in [26] for natural scene classification. LDA was also used in [45] for the discovery of object categories. The application of pLSA to scene classification was studied in [37], [7].

B. The Importance of Supervision

While the fundamental ideas of the following discussion apply to all topic models, we concentrate on LDA, which is the closest to the proposed architecture. In fact, the graphical model of Figure 8(b) *is* that of LDA. Figure 8(c) presents the complete version of this model, including the concept variable Y at the semantic level. This is one of the models proposed in [26]. Given the equivalence of the graphical models, it is worth discussing in detail the differences between the two approaches. The fundamental difference is the *level of abstraction* of the intermediate stage of the representation (topics vs. SMNs). While topics are learned in an unsupervised manner, SMN features have *explicit* semantics.

Recall the *semantic gap* between appearance features and visual classes. While text features (words) are intrinsically semantic, this is *not* the case for vision, where localized appearance features (e.g. edge segments) *have no semantic interpretation*. This is illustrated in Figure 9, where we present four groups of text (words) and appearance (edge segments) features *with identical distributions*. Because the word features are semantic, it is very difficult to construct a group (sentence) with the same words that is semantically far from the others. This is absolutely not the case for vision, where equivalence of feature distributions places almost no constraint on the group semantics. As the figure shows, the exact same segments can very easily be used to construct groups that depict completely unrelated concepts. The fact that *equivalence of feature distributions does not translate into semantic equivalence* is denoted a semantic gap.

While the semantic gap is small for text (semantic features), it is large for images. Thus, the success of a representation for text classification is an unreliable predictor of its success for image classification. In particular, the observation that unsupervised topic discovery produces semantic topics for text [6], [21], is very weak evidence that it will be successful for visual recognition. In fact, Figure 9 shows that it cannot. In the absence of explicit supervision for topic semantics, it is impossible to learn that the four edge groupings of (c) belong to different topics. On the contrary, the four groups form a perfect appearance cluster, since their segment histograms *are identical*. Unfortunately, due to the semantic gap, this cluster has no well defined semantics *as a whole*. Hence, unsupervised topic learning has no ability to bridge the semantic gap between local appearance and visual classes. This is unlike the proposed architecture, where SMN features are learned with explicit supervision, and it does make sense to talk about a *semantic space*.

It should be emphasized that in this toy example, although explicit topic supervision results in four classes of *identical* distribution (a highly suboptimal clustering under any unsupervised learning criteria), it produces the *semantically correct* statistical description of the data under the chosen image representation. Note that, under this model, all images of Figure 9(right) have an equal chance of being assigned to any of the classes. This is a classifier of higher probability of error than that learned without supervision. In fact, it is the weakest possible classifier. On the other hand, unsupervised topic modeling produces a much stronger classifier: all images assigned to one class with high probability, other classes mostly noise. In summary, the supervised model reflects *both* the true semantics of the data and the ambiguity of the image representation. It attempts to perform the *right* classification but can only do so with high uncertainty. The unsupervised model *invents* an alternative classification problem, which has nothing to do with the image semantics but *can be* solved very accurately. In addition to producing a semantically useless image description, it is also confident on its accuracy.

VII. Experimental Setup

In this section, we describe the experimental setup used to evaluate performance of the proposed contextual modeling.

A. Datasets

To test the proposed contextual modeling framework, we adopt datasets previously used in the scene classification and image retrieval literatures.

1) Scene Classification: Scene classification results are presented for two publicly available datasets.

Natural Scene Categories (N15, N13, N8) consists of fifteen categories (N15) of natural scenes, first proposed in [25]. This dataset was constructed using the 13 scene categories (N13) initially used by [26], [7]. The 13 scene categories themselves contain 8 categories (N8) originally used in [34], [37], [7]. This dataset allows direct comparison with published results on scene classification. Each category contains 200 to 400 images, of average size 270×250 pixels. 100 images per scene are used to learn the model, the remaining being used as test set. All experiments are repeated six times, with random train/test splits.

Corel Image Collection (C50, C43) consists of images from 50 Corel Stock Photo CDs, where each CD contains 100 images of a common scene. Each image in this dataset is also labeled with 1-5 semantic concepts. This annotated set is commonly used for the evaluation of image annotation systems [13], [15], [24]. We construct two different datasets from this collection. The first, referred to as C50, contains 50 scene classes, each corresponding to one CD in the collection. For each CD, 90 images are used to learn class models and the remaining for testing. It has been argued that CD labels lead to an easy classification problem [59] as there is high variability between images from different CDs and high similarity among those from the same CD. To address these concerns, we construct another dataset from this collection (C43) that uses a set of manual annotations (disjoint from the CD labels) as ground truth. 43 semantic concepts are chosen from the set of annotations of [13] (those with a minimum of 100 annotated images) and 100 images are randomly selected per concept. Since an image can be labeled with more than one concept, this results in a total of 3102 images. Of these, 2766 are randomly selected to create a test set with approximately 90 images per label, and the remainder are used for testing. A correct classification is declared whenever the top predicted label matches any of the groundtruth labels. All images were normalized to size 181×117 or 117×181 and converted from RGB to the YBR color space.

2) Image Retrieval: To evaluate retrieval performance, we use two datasets proposed in [39].

Corel Image Collection (C15) consists of $1,500$ images from another 15 Corel Stock Photo CDs, divided into a retrieval set of $1,200$ images and a query set of 300 images. CD themes are used as the ground truth image concepts, creating a 15-dimensional semantic space.

Flickr Images (F18) consists of $1,800$ images from `www.flickr.com`. These images are shot by flickr users, and hence differ from the Corel Stock Photos, which have been shot professionally. A set of $1,440$ images serves as the retrieval dataset, and the remaining 360 as the query set. Image annotations are those used in [39]. The semantic space is 18-dimensional.

Note that, for all datasets except C43, each image is explicitly annotated with just one concept, even though it may depict multiple. Thus, the co-occurrence information learned from these datasets is purely data driven. In C43, although multiple annotations are available per image, their co-occurrences are not explicitly used to learn context. In summary, no high level co-occurrence information is used to train the contextual models.

B. Appearance Features

Two feature transforms are used for appearance representation.

1) Scale Invariant Feature Transform: SIFT descriptors are computed at a set of feature points selected 1) by interest point detection (SIFT-INTR), or 2) on a dense regular grid (SIFT-GRID). Interest points are computed with three saliency measures — Harris-Laplace, Laplace-of-Gaussian, and Difference-of-Gaussian - and merged. These measures also provide scale information, which is used in the computation of SIFT descriptors. For a dense

TABLE II
IMPACT OF INFERENCE MODEL ON CLASSIFICATION ACCURACY.

	Classification Accuracy (%)		
Model	Appearance	Contextual	
		Image	RandomPatch
Figure 2, Eq (3)	71.67 ± 1.17	71.67 ± 1.17	-
Figure 8(a), Eq (10)	71.67 ± 1.17	**73.33** ± 0.69	**77.20** ± 0.39
Figure 8(b), Eq (12)	54.97 ± 0.58	**73.43** ± 0.99	75.14 ± 0.75

TABLE III
IMPACT OF APPEARANCE SPACE ON CLASSIFICATION ACCURACY.

Feature	Classification Accuracy (%)		Gain
	Appearance Models	Contextual Models	
SIFT-GRID using (10)	71.67 ± 1.17	**77.20 ± 0.39**	7.7%
SIFT-GRID using (12)	54.97 ± 0.58	**75.14 ± 0.75**	36.7%
SIFT-INTR	68.58 ± 0.41	**72.65 ± 0.56**	5.9%
DCT	47.33 ± 1.22	**73.05 ± 0.54**	54.3%

grid, feature points are sampled every 8 pixels. SIFT descriptors[1] are then computed over a 16×16 neighborhood around each feature point. The two strategies yield about 1000 samples per image.

2) Discrete Cosine Transform: DCT features are computed on a dense regular grid, with a step of 8 pixels. 8×8 image patches are extracted around each grid point, and 8×8 DCT coefficients computed per patch and color channel. For monochrome images this results in a feature space of 64 dimensions. For color images the space is 192 dimensional. In this case, appearance distributions are learned in the 129 dimensional subspace composed of the first 43 DCT coefficients from each channel. For datasets exclusively comprised of color images, only the DCT features are used.

VIII. RESULTS

A number of classification experiments were performed (N15 dataset) to evaluate the impact of the various parameters of the proposed contextual representation on recognition performance.

A. Designing the Semantic Space.

In Section V, we discussed three strategies to compute Image-SMNs. Table II reports their classification accuracy, for both appearance and contextual modeling with SIFT-GRID. Contextual models learned from SMNs computed with (3) fail to improve upon the (already high performing) appearance classifiers. This is not totally surprising, since these SMNs lack co-occurrence information (see discussion of Figure. 7). In comparison, SMNs computed with (10) or (12) are rich in such information, enabling contextual models to outperform their appearance counterparts.

Note that, although the LDA-like inference algorithm of (12) yields significantly lower classification performance at the appearance level than that of (10), both strategies attain a classification accuracy of $\sim$ 73.3% at the contextual level. Note also that, despite much weaker performance at appearance-level than (3), (12) performs substantially better at the contextual level. Together, these results suggest that the recognition performance at the appearance level is not necessarily a good predictor of performance at the contextual level. In particular, the relative performances of the three inference procedures advise against inference procedures that make hard decisions at the lower levels of recognition.

To increase the cardinality of the training sets used for contextual modeling, 800 random sets of 30 patches are sampled per image, yielding 800 patch-SMNs per image. Image-SMNs are then computed from these, with (10) or (12). Table II reports the benefits of this data augmentation, showing that performance improves in both cases. For (10) classification accuracy improves from 73.33% to 77.20%, for (12) from 73.43% to 75.14%. Since (12) involves an iterative procedure, which is more expensive than the closed form of (10), and has weaker performance, we use (10) in the remaining experiments.

B. Number of Mixture Components

Figure 10(a) presents the classification performance as a function of the number of contextual mixture components, for SIFT-GRID, SIFT-INTR and DCT features. In all cases, a single Dirichlet distribution is insufficient to model the semantic co-occurrences of N15. As the number of mixture components increases from 1 to 8, performance rises substantially for SIFT (e.g. from 72.58% to 76.13% for SIFT-GRID), and dramatically (from 55.93% to 70.48%) for the DCT. Above 8 components, the gain is moderate in all cases, with a maximum accuracy of 77.20% for SIFT-GRID and 73.05% for the DCT. Figure 11 shows the cluster centers learned with a four-component Dirichlet mixture using DCT features, for the "street" and "forest" classes. These cluster centers can be interpreted as the SMNs of the dominant co-occurrence patters learned for these classes. Two interesting observations can be made. First, the class mixtures indeed account for different co-occurrence patterns: in both cases the four cluster centers are quite distinct. Second, not all cluster centers assign high probability to the feature vector which is namesake of the class. In the "street" example, although one of the centers assigns high probability to the "street" concept, the remaining ones assign higher probability to alternative concepts, e.g. "tall building", "inside city", "highway" etc. than to "street" itself.

C. Choice of Appearance Features

Table III compares the classification performance of the three appearance representations. In all cases, the contextual models yield improved performance, with a gain of $7.7\%, 5.9\%$ and over 54% for SIFT-GRID, SIFT-INTR and DCT, respectively. Note that the contextual models achieve high performance (over 72%) for *all* appearance features. More interestingly, this performance is almost unaffected by that of the underlying appearance classification, in the sense that very large variations in the latter lead to relatively small differences in the former.

This hypothesis was studied in greater detail, by measuring how contextual-level performance depends on the "quality" of the appearance classification. The number of Gaussian components in the appearance models was the parameter adopted to control this "quality". Figure 10(b) and (c) shows that decreasing this parameter leads to a *substantial* degradation of appearance-level recognition, for both SIFT and DCT. Nevertheless, the performance of the contextual classifiers, built with these appearance classifiers, *does not change substantially*. On the contrary, the contextual classifiers assure a classification gain that *compensates* for the losses in appearance classification. For SIFT-GRID, this

[1] Computed with the implementation of LEAR http://lear.inrialpes.fr/people/dorko/downloads.html

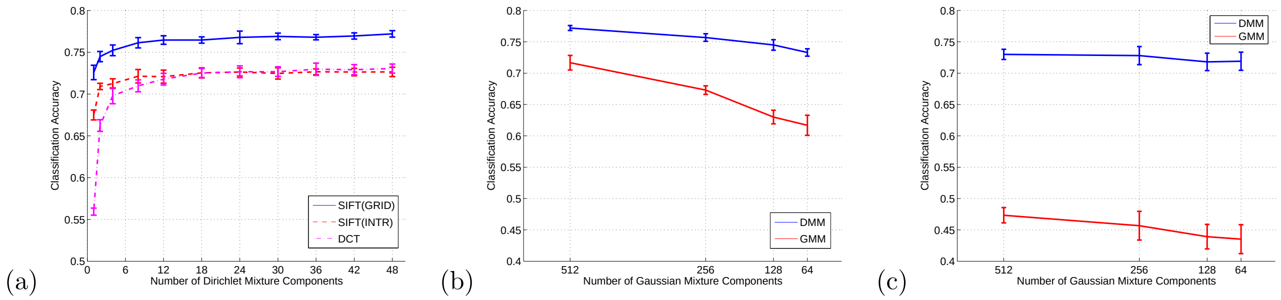


Fig. 10. (a) Classification accuracy as a function of the number of mixture components of the contextual class distributions, for both DCT and SIFT. (b) Dependence of appearance and contextual classification on the accuracy of the appearance modeling for SIFT-GRID features, (c) for DCT features. The performance of contextual classification remains fairly stable across the range of appearance models.

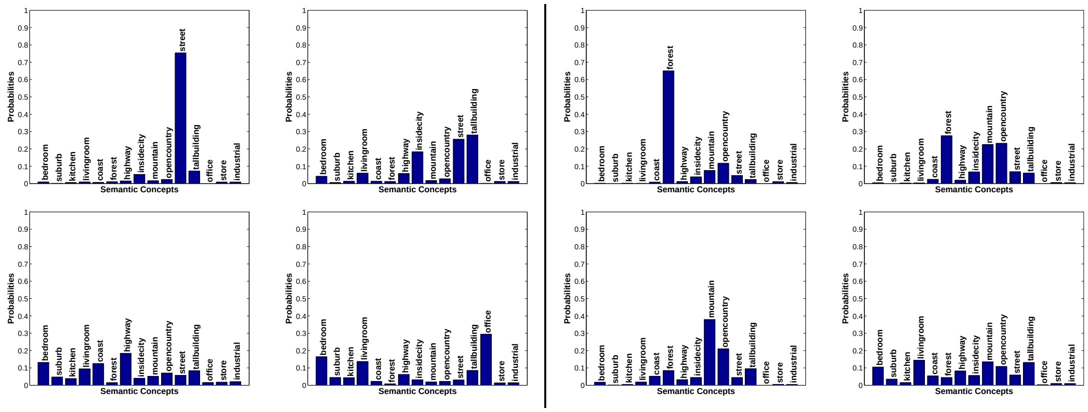


Fig. 11. Four cluster centers for the class “street” (left) and “forest” (right). Note that each class comprises different co-occurrence patterns.

gain ranges from about 20% at 64 Gaussian mixture components, to about 8% at 512. For the DCT, corresponding gains are of 65% and 54% respectively. In result, while the appearance classifier experiences a drop of 17% (21%) for DCT (SIFT-GRID) as the number of components is reduced from 512 to 64, the performance of contextual classification drops by only a small margin of 2% (5%).

Overall, the performance of the contextual classifier is not even strongly affected by the feature transformation adopted. While, at the appearance level, the performance of the DCT is not comparable to that of SIFT, the choice of transform is much less critical when contextual modeling is included: the two transforms lead to similar performance at the contextual level. This suggests that 1) any reasonable architecture could, in principle, be adopted for appearance classification, and 2) there is no need for extensive optimization at this level. This is an interesting conclusion, given that accurate appearance classification has been a central theme in the recognition literature over the last decades.

D. Some Examples

The ability of contextual modeling to compensate for classification noise at the appearance level can be observed by simple inspection of the posterior distributions at the two levels. Figure 12 shows two images from the “street” class of N15, and an image each from the “Ireland” and “Mayan ruins” CD of the Corel Collection. The SMN and CMN vectors computed from each image are shown in the second and third column, respectively. Two observations can be made. First, as discussed in Section. III-D, the SMN vectors can include substantial *contextual noise*, reflecting *both* types of concept co-occurrences. For example, patches from the first image (“street” class) have high probability under concepts such as “bedroom”, “livingroom”, “kitchen”, “inside city”, “tall building”. Some of these co-occurrences (“bedroom”, “livingroom”, “kitchen”) are due to patch ambiguities. Others (“inside city”, “tall building”) are consistent with the fact that the concepts are contextually dependent. The SMN representation has no power to disambiguate between the two types of co-occurrences. This is more pronounced for larger semantic spaces: the SMNs of Corel images (43 dimensional space) exhibit much denser co-occurrence patterns than those of N15.

Second, CMNs are remarkably noise-free for all semantic spaces considered. They capture the “gist” of the underlying scenes, assigning high probability only to truly contextual concepts. This increased robustness follows from the fact that contextual models learn the statistical structure of the contextual co-occurrences that characterize *all* SMNs associated with each class. This makes class models at contextual level mitigate ambiguity co-occurrences, which tend to be spurious, while accentuating true contextual co-occurrences, which are stable. Consider, for example, the top image in the fourth column. Its SMN is a frequently occurring training example for contextual models of “street”, “house”, “people” (this is true even though the image has low probability of “street” and “house” under appearance modeling), etc. On the other hand, it is an unlikely training pattern for contextual models of “bear” and “hills”, which only accidentally co-occur with “street” or “house”. Hence, this SMN has large posterior probability under contextual models for “house” and “street”, but not for “bear” or “hills”.

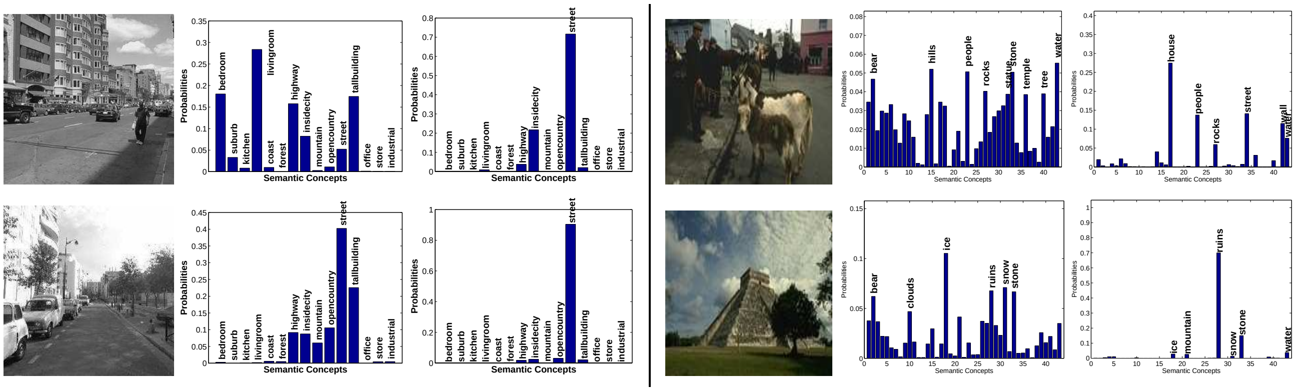


Fig. 12. left) Two images from the "street" class of N15, and right) an image each from the "Ireland" and "Mayan ruins" CD of the Corel collection. Also shown with the images are the SMN and CMN vectors (middle and right column respectively). Notice that the CMN vectors are noise-free and capture the "gist" of the image.

E. Complexity

In this section we report approximate running times for training and testing, under both the appearance and contextual class models. All experiments are conducted on an 2x Intel Xeon E5504 Quad-core 2.00GHz processor, with average image size of 270 × 250 pixels. Learning of appearance models requires computing SIFT/DCT features, which takes about 800/20ms per image respectively. Given these features, 512 component Gaussian mixture models are learned from 100 training images in about 3 minutes per class, using the hierarchical approach of [55]. For testing, computing the likelihood of a given image requires about 50ms per class. These likelihoods serve as features for the contextual models. A 42 component Dirichlet mixture model, learned from 100 training images, with 800 SMNs per image, requires about 2 minutes to learn. During testing, it takes about 30ms to compute the likelihood of an image under each contextual class model.

IX. Comparison with Previous Work

In this section we compare the proposed contextual recognition with existing solutions to scene classification and image retrieval.

A. Scene Classification

Given the posterior probabilities of (7), MPE scene classification can be implemented by application of Bayes rule. This consists of assigning image $\mathcal{I}$, of SMN π, to the scene class y of largest posterior $P_{Y|\Pi}(y|\pi)$. Table IV compares the resulting classification accuracies for N15, N13, and N8, with those of many methods in the literature. A number of observations can be made from the table. First, contextual modeling achieves the best results on all three datasets. Its performance is quite superior to that of topic discovery models (LDA [26], pLSA [7], [37]), of which only [7] is remotely competitive. Even so, the classification rates of the latter (72.7% on N15 , 74.7% on N13, and 82.5% on N8) are well below those of the former (77.2%, 80.86%, and 85.6%). Somewhat closer to this (74.8% on N15, 74.7% on N13) is the performance of SVMs with the bag-of-words representation (BoW)[2]. Note, however, that these require much higher dimensional spaces, e.g. a 400 visual-word vocabulary [25], and storage of a number of support vectors that grows with the number of classes and training examples. Contextual modeling has lower dimensionality, lower complexity, and achieves a higher classification accuracy[3]. Also reported is a baseline with discriminative learning [40] where an SVM classifier is applied to the vector of outputs of the appearance classifiers. Again, the proposed context models achieve superior classification performance on all datasets.

Within the area of context modeling, e.g. comparing to the methods of [34], [27], the proposed approach is again more effective. For the N8 (N13, N15) dataset, [34] ([22], [25]) report a classification accuracy of 83.7% (55%, 45.3%[4]), respectively, using the "gist" features of [34]. The corresponding figures for the proposed contextual models are 85.6% (80.86%, 77.2%).

Finally, Table V presents classification results for the C50 and C43 datasets. Contextual modeling again improves on the classification accuracy achievable with appearance classifiers. For C50 the absolute gain is of 4.2%, for C43 of 3%. When compared to the top performing published methods on the natural scene dataset [25], [7] the proposed contextual modeling again achieves significantly higher accuracy. On C50, its accuracy is 57.8% while [25] and [7] achieve classification rates of 48.4% and 40.2%, respectively. On C43, the corresponding numbers are 42.9%, 36.3%, and 33.0%. Overall, it can be concluded that the proposed contextual modeling consistently outperforms existing context-based scene classification methods in the literature.

B. Image Retrieval Performance

Finally, the benefits of holistic context modeling were evaluated on the task of content based image retrieval, using the well known

[2] This representation is obtained by vector quantizing the space of descriptors and representing an image with a visual word histogram.

[3] We note that better results have been reported for an extension of the BoW representation that includes a weak encoding of spatial information [25], [62]. These results are the current state-of-the-art for N15: 81.4% [25] using a SVM classifier on an 8400 dimensional space; 85.2% [62] using a nearest neighbor classifier on an 8192 dimensional space. Note that the performance of these approaches without the additional spatial encoding is 74.8% and 75.8%, respectively, which is well below the 77.2% achieved by the proposed contextual models. Although contextual classification could also be augmented with weak encoding of spatial information — one possibility is to learn contextual class models for different image sub-regions and model the overall contextual class model as a mixture of these sub-region models — it remains to be determined if the gains would be as large as for the BoW representation. We leave this as a topic for future work.

[4] Using a 16 dimensional "gist" like feature instead of the commonly used 512 dimensions.

TABLE IV
CLASSIFICATION RESULTS ON NATURAL SCENE CATEGORIES.

Method	Classif.	Dims.[a]	Accuracy (%)
	N15 Dataset		
Contextual Models	Bayes	**15**	**77.20 ± 0.39**
pLSA [7][b]	SVM	40	72.7
pLSA [25]	SVM	60	63.3
LDA [26][e]	Bayesian	40	59.0
“gist” like [25]	SVM	16	45.3 ± 0.5
BoW [25]	SVM	400	74.8 ± 0.3
BoW [25]	SVM	200	72.2 ± 0.6
Bag of Concepts [28][c]	SVM	100	73.01
Kernel Codebook [52]	SVM	3200	∼75[d]
Diffusion Distance [29]	SVM	2000	74.9
SIS [11]	SVM	200	74.94
Semantic Space [40]	SVM	15	73.95 ± 0.74
	N13 Dataset		
Contextual Models	Bayes	**13**	**80.86 ± 0.50**
LDA [26]	Bayesian	40	65.2
pLSA [7][b]	SVM	35	74.3
pLSA [37]	SVM	40	60.8
pLSA [25]	SVM	60	65.9
BoW [25]	SVM	200	74.7
Taxonomy [3]	Bayesian	40	68
“gist” features [22]	SVM	512	∼55[d]
Semantic Space [40]	SVM	13	77.57 ± 1.12
	N8 Dataset		
Contextual Models	Bayes	**8**	**85.60 ± 0.70**
Context Ancestry [27]	Logistic	484	82
pLSA [7][b]	SVM	25	82.5
HDP-HMT [23]	Bayesian	200	84.5
“gist” [34][f]	SVM	512	83.7
Semantic Space [40]	SVM	8	84.24 ± 0.71

[a] Dimensionality of the space on which classification is performed
[b] Uses half of the dataset for training
[c] Uses a subset of test images per concept
[d] Accuracy estimated from figure
[e] Our implementation of the algorithm
[f] Gist features implicitly uses weak spatial information

TABLE V
CLASSIFICATION RESULTS ON COREL COLLECTION.

Method[a]	Classif.	Dims.	Accuracy (%)
	C50 Dataset		
Contextual Models	Bayes	**50**	**57.8**
Appearance Models	Bayes	129	53.6
Bag of Words [25]	SVM	512	48.4
pLSA [7]	SVM	50	40.2
LDA [26]	Bayes	50	31.0
	C43 Dataset		
Contextual Models	Bayes	**43**	**42.9**
Appearance Models	Bayes	129	39.9
Bag of Words [25]	SVM	512	36.3
pLSA [7]	SVM	50	33.0
LDA [26]	Bayes	50	24.6

[a] Our implementation of the algorithms

query-by-example paradigm. This is a nearest-neighbor classifier, where a vector of global image features extracted from a query image is used to retrieve the images of closest feature vector in an image database. In previous work [39], we have shown that state-of-the-art results for this type of operation are obtained by using appearance-level posterior distributions (SMNs) as feature vectors. In this work, we compare results of using the distributions obtained at the contextual (CMN) and appearance (SMN) levels. The similarity between the distributions of the query and database images is measured with the Kullback-Leibler divergence [39].

Figure 13(left), presents precision-recall (PR) curves on C15 and F18. Also shown are the performance of the image matching system of [54], which is based on the MPE retrieval principle now used but does not rely on semantic modeling, and chance-level retrieval. Note how the precision of contextual modeling is *significantly* superior to those of the other methods at *all* levels of recall. For example, on C15, the mean-average precision (area under PR curve) of CMN (0.73) is 32% higher than that of SMN (0.55). The respective figures for F18 are 0.54 and 0.39, a gain of over 38%. Overall, the PR curves of CMN are remarkably flat, attaining high precision at high levels of recall. This is unlike any other retrieval method that we are aware of. It indicates very good generalization: while most retrieval approaches (even image matching) can usually find a few images in the class of the query, it is much more difficult to generalize to images in the class that *are not* visually similar to the query.

Figure 13(right) illustrates the improved generalization of contextual modeling. It presents retrieval results for the three systems (top three rows of every query show the top retrieved images using visual matching, SMN, and CMN respectively). The first column shows the queries while the remaining columns show the top five retrieved images. Note how visual matching has no ability to bridge the semantic gap, simply matching semantically unrelated images of similar color and texture. This is unlike the semantic representations (SMN and CMN) which are much more effective at bridging the gap, leading to a much smaller number of semantically irrelevant matches. In particular, the ability of the CMN-based system to retrieve images in the query's class is quite impressive, given the high variability of visual appearance.

X. CONCLUSION

In this work, we have proposed an approach to context modeling based on the probability of co-occurrence of objects and scenes. The proposed modeling is quite simple, and builds upon the availability of robust appearance classifiers. Images are represented by their posterior probabilities with respect to a set of contextual models, built upon the bag-of-features image representation through two layers of probabilistic modeling. The first layer represents the image in a semantic space, where each dimension encodes an appearance-based posterior probability with respect to a visual concept. This representation has a higher level of abstraction than bag-of-features but suffers from a certain amount of contextual noise, due to the inherent ambiguity of classifying image patches. The second layer enables robust inference in the presence of this noise, by modeling the distribution of each concept in the semantic space. The image is then represented by its posterior probabilities with respect to these distributions. This was shown to produce posterior distributions that emphasize concept co-occurrences due to true contextual relationships and inhibit accidental co-occurrences due to ambiguity.

The overall representation is similar to a topic model, but where topics are learned in a supervised manner. Supervised learning is a necessary condition for overcoming the semantic gap between the low-level patch representation and the higher-level contextual relationships. While multiple instance learning is required to cope with the uncertainty of the appearance representation, multiple instance inference was shown ineffective. Best results are obtained with weaker, patch-based, inference that leads to an appearance representation of higher entropy. This

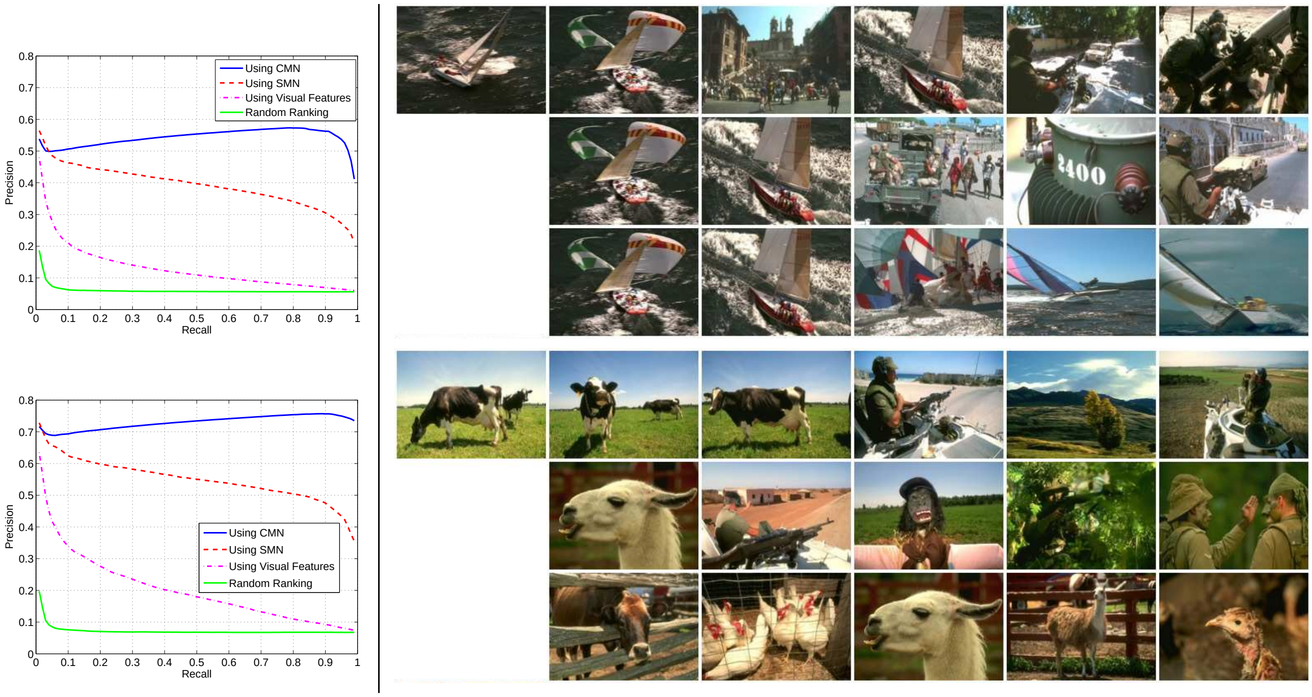


Fig. 13. right) Precision-recall curves achieved with SMN, CMN, visual matching and chance level image retrieval for (top) F18, (bottom) C15 datasets. left) Retrieval results for four image queries shown on the left-most column. The first, second, and third row of every query show the five top matches using image matching, SMN, and CMN-based retrieval, respectively.

prevents a greedy commitment to premature image explanations that, while consistent with appearance statistics, do not take context into account. The latter goal is better served by inference procedures that assign non-zero probability to multiple plausible classes, at the appearance level. Interestingly, we found a weak correlation between the quality of the appearance classification and the corresponding quality at the contextual level. In fact, some variations of the representation with weak appearance-level performance were top-performers at the contextual level. It appears that, while supervision is critical to bridging the semantic gap during learning, soft appearance-level decisions are more effective during inference. This is an interesting finding, given the emphasis on highly accurate appearance classification in the literature.

The contextual representation was shown to outperform the appearance representation in the tasks of scene classification and image retrieval. In both cases, it was also shown that, despite its simplicity, the proposed contextual models are superior to various previous proposals in the literature. The gains with respect to appearance modeling were shown to hold irrespectively of the choice and accuracy of the underlying appearance models.

References

[1] M. Bar. Visual objects in context. *Nature Reviews Neuroscience*, 5(8):617–629, 2004.

[2] K. Barnard, P. Duygulu, N. de Freitas, D. Forsyth, D. Blei, and M. I. Jordan. Matching words and pictures. *Journal of Machine Learning*, 3:1107–1135, 2003.

[3] E. Bart, I. Porteous, P. Perona, and M. Welling. Unsupervised learning of visual taxonomies. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2008.

[4] I. Biederman, R. Mezzanotte, and J. Rabinowitz. Scene perception: detecting and judging objects undergoing relational violations. In *Cognitive Psychology*, volume 14, pages 143–77, 1982.

[5] D. Blei and M. Jordan. Modeling annotated data. In *Proc. ACM SIGIR conf. on Research and development in information retrieval*, 2003.

[6] D. Blei, A. Ng, and M. Jordan. Latent dirichlet allocation. *The Journal of Machine Learning Research*, 3:993–1022, 2003.

[7] A. Bosch, A. Zisserman, and X. Munoz. Scene classification using a hybrid generative/discriminative approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(4):712, 2008.

[8] D. Cai, X. He, and J. Han. Efficient kernel discriminant analysis via spectral regression. *Proc. Int. Conf. on Data Mining*, 2007.

[9] G. Carneiro, A. Chan, P. Moreno, and N. Vasconcelos. Supervised learning of semantic classes for image annotation and retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(3):394–410, March, 2007.

[10] A. Chan and N. Vasconcelos. Probabilistic kernels for the classification of auto-regressive visual processes. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, 2005.

[11] H. Cheng, Z. Liu, and J. Yang. Sparsity induced similarity measure for label propagation. *IEEE International Conference on Computer Vision*, 2009.

[12] R. Datta, D. Joshi, J. Li, and J. Wang. Image retrieval: Ideas, influences, and trends of the new age. *ACM Computing Surveys*, 39:65, 2007.

[13] P. Duygulu, K. Barnard, N. Freitas, and D. Forsyth. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In *European Conference on Computer Vision*, 2002.

[14] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2009.

[15] S. Feng, R. Manmatha, and V. Lavrenko. Multiple bernoulli relevance models for image and video annotation. In *IEEE Conference on Computer Vision and Pattern Recognition*, Washington DC, 2004.

[16] R. Fergus, P. Perona, and A. Zisserman. Object class recognition by unsupervised scale-invariant learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, pages 264–271, June 2003.

[17] M. Fink and P. Perona. Mutual boosting for contextual inference. *Neural Information Processing Systems*, 2004.

[18] C. Galleguillos, A. Rabinovich, and S. Belongie. Object categorization using co-ocurrence, location and appearance. *IEEE Conference on Computer Vision and Pattern Recognition*, 2008.

[19] K. Grauman and T. Darrell. The pyramid match kernel: Efficient learning with sets of features. *Journal of Machine Learning Research*, 8:725–760, 2007.

[20] G. Heitz and D. Koller. Learning spatial context: Using stuff to find things. *European Conference on Computer Vision, Marseille, France*, page 30, 2008.

[21] T. Hofmann. Probabilistic latent semantic indexing. *Proc. ACM SIGIR conf. on Research and development in information retrieval*, 1999.

[22] A. Joshi, F. Porikli, and N. Papanikolopoulos. Multi-class active learning for image classification. In *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2009.

[23] J. Kivinen, E. Sudderth, and M. Jordan. Learning multiscale representations of natural scenes using dirichlet processes. In *IEEE International Conference on Computer Vision*. Citeseer, 2007.

[24] V. Lavrenko, R. Manmatha, and J. Jeon. A model for learning the semantics of pictures. In *Advances in Neural Information Processing Systems*, 2003.

[25] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. *IEEE Conference on Computer Vision and Pattern Recognition*, 2005.

[26] F.-F. Li and P. Perona. A bayesian hierarchical model for learning natural scene categories. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 524–531, 2005.

[27] J. Lim, P. Arbeláez, C. Gu, and J. Malik. Context by region ancestry. In *IEEE International Conference on Computer Vision*. Citeseer, 2010.

[28] J. Liu and M. Shah. Scene modeling using co-clustering. *IEEE International Conference on Computer Vision*, 2007.

[29] J. Liu, Y. Yang, and M. Shah. Learning semantic visual vocabularies using diffusion distance. In *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2009.

[30] D. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[31] D. MacKay. *Information theory, inference, and learning algorithms*. Cambridge Univ Pr, 2003.

[32] J. Magalhães, S. Overell, and S. Rüger. A semantic vector space for query by image example. *Proc. ACM SIGIR conf. on Research and development in information retrieval*, 2007.

[33] A. Oliva and P. Schyns. Diagnostic colors mediate scene recognition. *Cognitive Psychology*, 41(2):176–210, 2000.

[34] 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.

[35] A. Oliva and A. Torralba. Building the gist of a scene: The role of global image features in recognition. *Visual Perception*, 2006.

[36] J. Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in Large Margin Classifiers*, 6174, 1999.

[37] P. Quelhas, F. Monay, J. Odobez, D. Gatica-Perez, and T. Tuytelaars. A thousand words in a scene. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(9):1575–1589, 2007.

[38] A. Rabinovich, A. Vedaldi, C. Galleguillos, E. Wiewiora, and S. Belongie. Objects in context. *IEEE International Conference on Computer Vision*, pages 1–8, 2007.

[39] N. Rasiwasia, P. Moreno, and N. Vasconcelos. Bridging the gap: Query by semantic example. *IEEE Transactions on Multimedia*, 2007.

[40] N. Rasiwasia and N. Vasconcelos. Scene classification with low-dimensional semantic spaces and weak supervision. *IEEE Conference on Computer Vision and Pattern Recognition*, 2008.

[41] N. Rasiwasia and N. Vasconcelos. Holistic context modeling using semantic co-occurrences. *IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

[42] L. Renninger and J. Malik. When is scene identification just texture recognition? *Vision Research*, 44(19):2301–2311, 2004.

[43] H. Rowley, S. Baluja, and T. Kanade. Neural network-based face detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 203–208, 1996.

[44] J. Shotton, J. Winn, C. Rother, and A. Criminisi. Textonboost for image understanding: Multi-class object recognition and segmentation by jointly modeling appearance, shape and context. *International Journal of Computer Vision*, pages 1–22, 2007.

[45] J. Sivic, B. Russell, A. Efros, A. Zisserman, and W. Freeman. Discovering object categories in image collections. *IEEE International Conference on Computer Vision*, 1:65, 2005.

[46] A. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain. Content-based image retrieval: the end of the early years. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(12):1349–1380, 2000.

[47] J. Smith, M. Naphade, and A. Natsev. Multimedia semantic indexing using model vectors. *IEEE International Conference on Multimedia and Expo*, pages 445–448, 2003.

[48] E. Sudderth, A. Torralba, W. Freeman, and A. Willsky. Learning hierarchical models of scenes, objects, and parts. *IEEE International Conference on Computer Vision*, 2, 2005.

[49] A. Torralba. Contextual priming for object detection. *International Journal of Computer Vision*, pages 169–191, 2003.

[50] A. Torralba, K. Murphy, and W. Freeman. Contextual models for object detection using boosted random fields. *Advances in Neural Information Processing Systems*, 2004.

[51] L. Torresani, M. Szummer, and A. Fitzgibbon. Efficient object category recognition using classemes. *ECCV*, pages 776–789, 2010.

[52] J. van Gemert, J. Geusebroek, C. Veenman, and A. Smeulders. Kernel codebooks for scene categorization. *European Conference on Computer Vision*, pages 696–709, 2008.

[53] M. Vasconcelos, N. Vasconcelos, and G. Carneiro. Weakly supervised top-down image segmentation. *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1001–1006, 2006.

[54] N. Vasconcelos. Minimum probability of error image retrieval. *IEEE Transactions on Signal Processing*, August 2004.

[55] N. Vasconcelos. Image indexing with mixture hierarchies. In *IEEE Conference on Computer Vision and Pattern Recognition*, Kawai, Hawaii, 2001.

[56] P. Viola and M. Jones. Robust real-time object detection. *International Journal of Computer Vision*, 1(2), 2002.

[57] J. Vogel and B. Schiele. A semantic typicality measure for natural scene categorization. *DAGM04 Annual Pattern Recognition Symposium*, 2004.

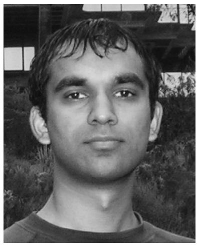
[58] G. Wang, D. Hoiem, and D. Forsyth. Learning image similarity from flickr groups using stochastic intersection kernel machines. In *IEEE International Conference on Computer Vision*, pages 428–435, 2009.

[59] T. Westerveld and A. de Vries. Experimental evaluation of a generative probabilistic image retrieval model on easy data. In *ACM SIGIR conf. on Research and development in information retrieval*. Citeseer, 2003.

[60] L. Wolf and S. Bileschi. A critical view of context. *International Journal of Computer Vision*, 69(2):251–261, 2006.

[61] H. Zhang, A. Berg, M. Maire, and J. Malik. Svm-knn: Discriminative nearest neighbor classification for visual category recognition. *IEEE Conference on Computer Vision and Pattern Recognition*, 2006.

[62] X. Zhou, N. Cui, Z. Li, F. Liang, and T. Huang. Hierarchical gaussianization for image classification. In *IEEE 12th International Conference on Computer Vision*, pages 1971–1977. IEEE, 2009.



Nikhil Rasiwasia Nikhil Rasiwasia received his B.Tech degree in electrical engineering form Indian Institute of Technology Kanpur (India) in 2005. He is a PhD student in the Statistical Visual Computing Lab in the ECE Department at the University of California, San Diego. His research interests are in the areas of computer vision and machine learning. He was recognized as an 'Emerging Leader in Multimedia' in 2008 by IBM T. J. Watson Research.

Nuno Vasconcelos Nuno Vasconcelos received the licenciatura in electrical engineering and computer science from the Universidade do Porto, Portugal, in 1988, and the MS and PhD degrees from the Massachusetts Institute of Technology in 1993 and 2000, respectively. From 2000 to 2002, he was a member of the research staff at the Compaq Cambridge Research Laboratory, which in 2002 became the HP Cambridge Research Laboratory. In 2003, he joined the Electrical and Computer Engineering Department at the University of California, San Diego, where he heads the Statistical Visual Computing Laboratory. He is the recipient of a US National Science Foundation CAREER award, a Hellman Fellowship, and has authored more than 50 peer-reviewed publications. His work spans various areas, including computer vision, machine learning, signal processing and compression, and multimedia systems. He is a member of the IEEE.