Chapter 6

Holistic Context Modeling

In this chapter we discuss some of the drawbacks of the proposed semantic image representation and introduce the framework of "holistic context modeling" that, while addressing these drawbacks, yields robust visual recognition systems.

6.1 Introduction

Recent 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 [10]. 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) [10].

In this chapter, 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 semantic representation of the images introduced in Chapter 2. Semantic image representation itself builds upon the *bag-of-features* (BoF) representation (see Chapter 2 for details regarding BoF representation), thereby inheriting several of its benefits. 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 BoF is a higher level of *abstraction*, which can lead to substantially better generalization — as established in Chapter 3, by comparing the performance of nearest-neighbors classification in an image retrieval context. However, the semantic representation also has some limitations that can be traced back to the BoF representation itself. 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 BoF 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 6.1).

Thus, 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. "bedroom" and "kitchen" in Figure 6.1) 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 chapter is organized as follows. Section 6.2 briefly reviews the literature on context modeling. Section 6.3 then discusses the limitations of semantic image representation built upon appearance classifiers and introduces contextual models. An extensive experimental evaluation of contextual modeling is then presented in Section 6.4, Section 6.5, and Section 6.6.

6.2 Related Work on 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 [115, 43], to explicit learning of the spatial relationships between objects [47, 174], or an object and its neighboring image regions [57]. 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” [104, 166, 77, 105, 74]. Various recent works have shown that semantic descriptions of natural images can be obtained with these representations, without explicit image segmentation [104]. This is consistent with evidence from the psychology [103] and cognitive neuroscience [3] 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, [104] models scenes according to the differential regularities of their second order statistics. A second approach is to rely on the BoF/BoW representation. Here, low-level features are computed locally and aggregated across the image, to form a holistic context model [166, 77, 121]. 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 [104, 74].

The proposed context modeling 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 [77, 114, 17]. These models were originally proposed in the text literature, to address the ambiguity of BoW. 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) [14, 58]. 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 [12, 4, 129, 140, 77, 114, 17]. 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 [83, 114, 74].

This issue is discussed in detail in the next chapter, where we argue that unsupervised topic discovery is not a good idea for recognition and 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 [120, 169, 147]. In these works a classifier is first learned for a given keyword vocabulary — [169, 147] learn discriminative classifiers from `flickr/bing` images, [120] 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.

6.3 Semantics-based Models and Context Multinomials

6.3.1 Limitations of Semantic Representations

One major source of difficulties is that semantic models built upon the BoF 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 BoF. 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 6.1. First, image patches frequently have ambiguous interpretation. When considered in isolation, they can be compatible with

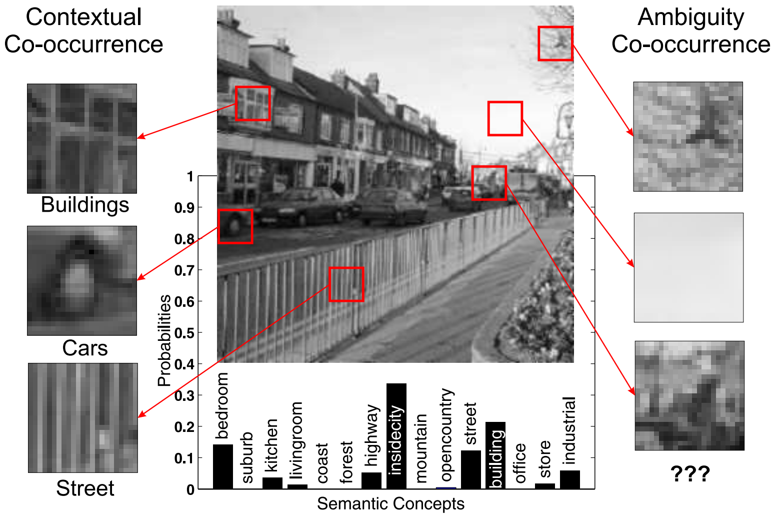


Figure 6.1: An image from the "street" class of the N15 dataset (See 6.4.1) along with its SMN. Also highlighted are the two notions of *co-occurrence*. *Ambiguity co-occurrences* on the right: image patches compatible with multiple unrelated classes. *Contextual co-occurrences* on the left: patches of multiple other classes related to "street".

many concepts. For example it is unclear that even a human could confidently assign the patches shown on the right of Figure 6.1 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 6.1. 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*.

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 [42, 40] or object relationships [146, 47]. However, the introduction of such constraints increases complexity, and reduces the invariance of the representation, sacrificing generalization. A more robust alternative is to keep BoF, 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.

6.3.2 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 6.1 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 6.2, each concept k is modeled

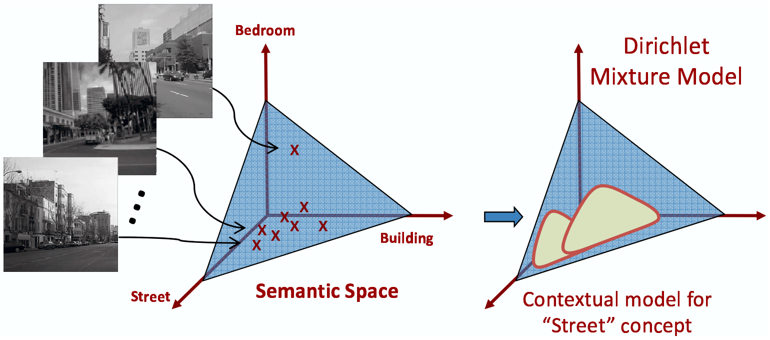


Figure 6.2: Learning the contextual model for the "street" concept, (6.1), on semantic space $\mathcal{S}$, from the set of all training images annotated with "street".

by the probability distribution of the SMNs derived from all training images in its training set, $\mathcal{D}_k$. We refer to this SMN distribution as the *contextual model* for k. If $\mathcal{D}_k$ is large and diverse, this model is dominated by the stable properties of the features drawn from concept k. In this case, the features are SMNs and their stable properties are the true contextual relationships of k. Hence, concept models assign high probability to regions of the semantic space occupied 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 Chapter 2, 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*.

6.3.3 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 K defined on the index set $k \in \{1, \ldots, K\}$ 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, K 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 k in $\mathcal{S}$ is the mixture of Dirichlet distributions

$$P_{\mathbf{\Pi}|K}(\boldsymbol{\pi}|k; \Lambda^k) = \sum_m \beta_m^k \mathcal{D}ir(\boldsymbol{\pi}; \boldsymbol{\alpha}_m^k). \tag{6.1}$$

This model has parameters $\Lambda^k = \{\beta_m^k, \boldsymbol{\alpha}_m^k\}$, where β_m is a probability mass function ($\sum_m \beta_m^k = 1$). $\mathcal{D}ir(\boldsymbol{\pi}; \boldsymbol{\alpha})$ a Dirichlet distribution of parameter $\boldsymbol{\alpha} = \{\alpha_1, \ldots, \alpha_L\}$,

$$\mathcal{D}ir(\boldsymbol{\pi}; \boldsymbol{\alpha}) = \frac{\Gamma(\sum_{i=1}^L \alpha_i)}{\prod_{i=1}^L \Gamma(\alpha_i)} \prod_{i=1}^L (\pi_i)^{\alpha_i - 1} \tag{6.2}$$

and $\Gamma(.)$ the Gamma function. As illustrated in Figure 6.2, the parameters Λ^k are learned from the SMNs $\boldsymbol{\pi}$ of all images in $\mathcal{D}_k$, i.e. the images annotated with the k^{th} concept in $\mathcal{L}$. Learning is implemented by maximum likelihood estimation, using the generalized expectation-maximization (GEM) algorithm discussed in Appendix B.

Figure 6.3 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

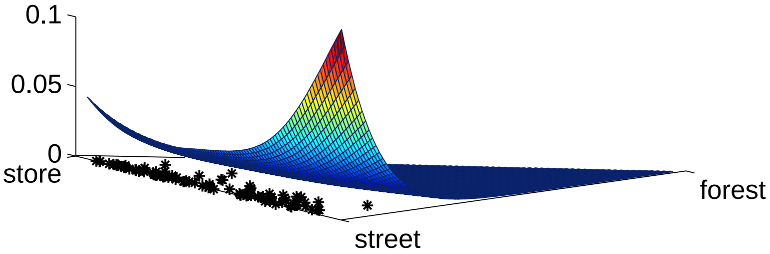


Figure 6.3: 3-component Dirichlet mixture learned for the concept "street". Also shown, as "*", are the SMNs associated with each image. The Dirichlet mixture assigns high probability to the concepts "street" and "store".

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.

6.3.4 Contextual Space

The contextual models $P_{\Pi|K}(\boldsymbol{\pi}|k)$ 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

$$P_{K|\Pi}(k|\boldsymbol{\pi}) = \frac{P_{\Pi|K}(\boldsymbol{\pi}|k)P_K(k)}{P_{\Pi}(\boldsymbol{\pi})}. \tag{6.3}$$

This is the semantic space equivalent of (2.8) and, once again, we assume a uniform concept prior $P_K(k)$.

As in Chapter 2, it is also possible to design a new semantic space, by retaining all posterior contextual concept probabilities $\theta_k = P_{K|\Pi}(k|\boldsymbol{\pi})$. We denote the vector $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K)^T$ as the *contextual multinomial* (CMN) distribution of image $\mathcal{I}$. As illustrated in Figure 6.4, 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 6.4 we show that CMNs are much more reliable contextual descriptors than SMNs.

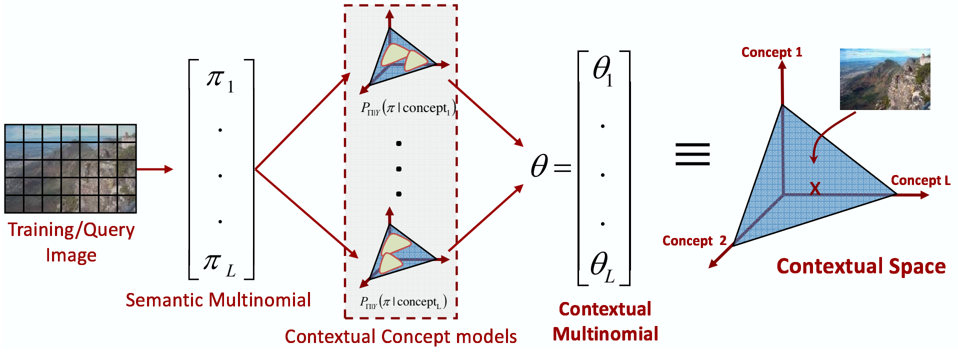


Figure 6.4: The Contextual multinomial (CMN) of an image as the vector of co-occurrence probabilities of contextually related concepts.

6.3.5 Data Augmentation

It should be noted that, similar to learning the semantic representation, 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. However, 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 labeled 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, $\boldsymbol{\pi}^{(n)}$ (see Section 2.3), 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 6.5.1, we show that this data augmentation strategy leads to significant improvements in classification accuracy.

6.4 Experimental Setup

In this section, we describe the experimental setup used to evaluate performance of the proposed contextual modeling. The evaluation consists of two vision tasks, viz. scene classification and image retrieval.

6.4.1 Datasets

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

Scene Classification

Scene classification results are presented for two publicly available datasets viz. "Natural Scene Categories" and "Corel Image Collection".

Natural Scene Categories (N15, N13, N8) We present results on all three subsets of the "Natural Scene Categories" dataset, viz. Natural15 (N15), Natural13 (N13) and Natural8 (N8). These dataset allows direct comparison with published results on scene classification. To learn the concept models, 100 images per scene are used, the remaining being used as test set. All experiments are repeated six times, with random train/test splits. A detailed description of these datasets are provided in Appendix. A.1.1.

Corel Image Collection (C50, C43) We also present results of the "Corel Image Collection" which has much higher number of classes as compared to the "Natural Scene Categories" dataset. We construct two different datasets from this collection, viz. Corel50 (C50) and Corel43(C43) with 50 and 43 classes respectively. For C50, 90 images from each CD are used to learn class models and the remaining for testing. For C43, 90 images per label are used to learn the class models and the remainder are used for testing. All images were normalized to size 181×117 or 117×181 and converted from RGB to the YBR color space. A detailed description of these datasets are provided in Appendix. A.1.3.

Image Retrieval

To evaluate retrieval performance, we use two datasets introduced in [119]. **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 and contextual space. A detailed description of C15 is provided in Appendix. A.1.3.

Flickr Images (F18) consists of $1,800$ images from `www.flickr.com` divided into 18 classes resulting in an 18 dimensional semantic and contextual space. A set of $1,440$ images serves as the retrieval dataset, and the remaining 360 as the query set. A detailed description of F18 is provided in Appendix. A.1.4.

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.

6.4.2 Appearance Features

Both SIFT and DCT features are used for appearance representation. SIFT features are computed either by interest point detection, SIFT-INTR, or on a dense regular grid SIFT-GRID. The two strategies yield about 1000 samples per image. 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

Table 6.1: Impact of inference model on classification accuracy.

	Classification Accuracy (%)		
Model	Appearance	Contextual	
		Image	RandomPatch
Figure 2.1, Eq (2.8)	71.67 ± 1.17	71.67 ± 1.17	-
Figure 2.5(a), Eq (2.21)	71.67 ± 1.17	**73.33** ± 0.69	**77.20** ± 0.39
Figure 2.5(b), Eq (2.23)	54.97 ± 0.58	**73.43** ± 0.99	75.14 ± 0.75

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.

6.5 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.

6.5.1 Designing the Semantic Space.

In Section 2.3, we discussed three strategies to compute Image-SMNs. 6.1 reports their classification accuracy, for both appearance and contextual modeling with SIFT-GRID. Contextual models learned from SMNs computed with (2.8) 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 Section 2.3). In comparison, SMNs computed with (2.21) or (2.23) are rich in such information, enabling contextual models to outperform their appearance counterparts.

Note that, although the LDA-like inference algorithm of (2.23) yields significantly lower classification performance at the appearance level than that of (2.21), 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 (2.8), (2.23) 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 (2.21) or (2.23). 6.1 reports the benefits of this data augmentation, showing that performance improves in both cases. For (2.21) classification accuracy improves from 73.33% to 77.20%, for (2.23) from 73.43% to 75.14%. Since (2.23) involves an iterative procedure, which is more expensive than the closed form of (2.21), and has weaker performance, we use (2.21) in the remaining experiments.

6.5.2 Number of Mixture Components

Figure 6.5(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 6.6 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

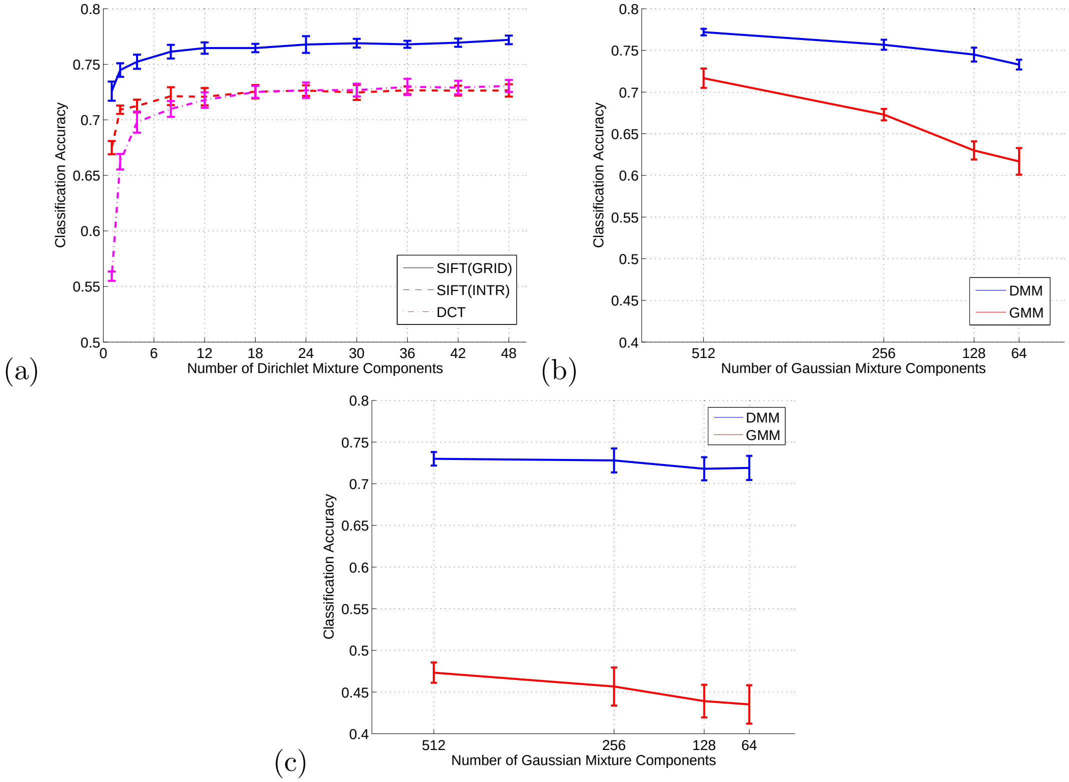


Figure 6.5: (a) Classification accuracy as a function of the number of mixture components of the contextual class distributions, for both DCT and SIFT. (b) Dependence of appearance and contextual classification on the accuracy of the appearance modeling for SIFT-GRID features, (c) for DCT features. The performance of contextual classification remains fairly stable across the range of appearance models.

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.

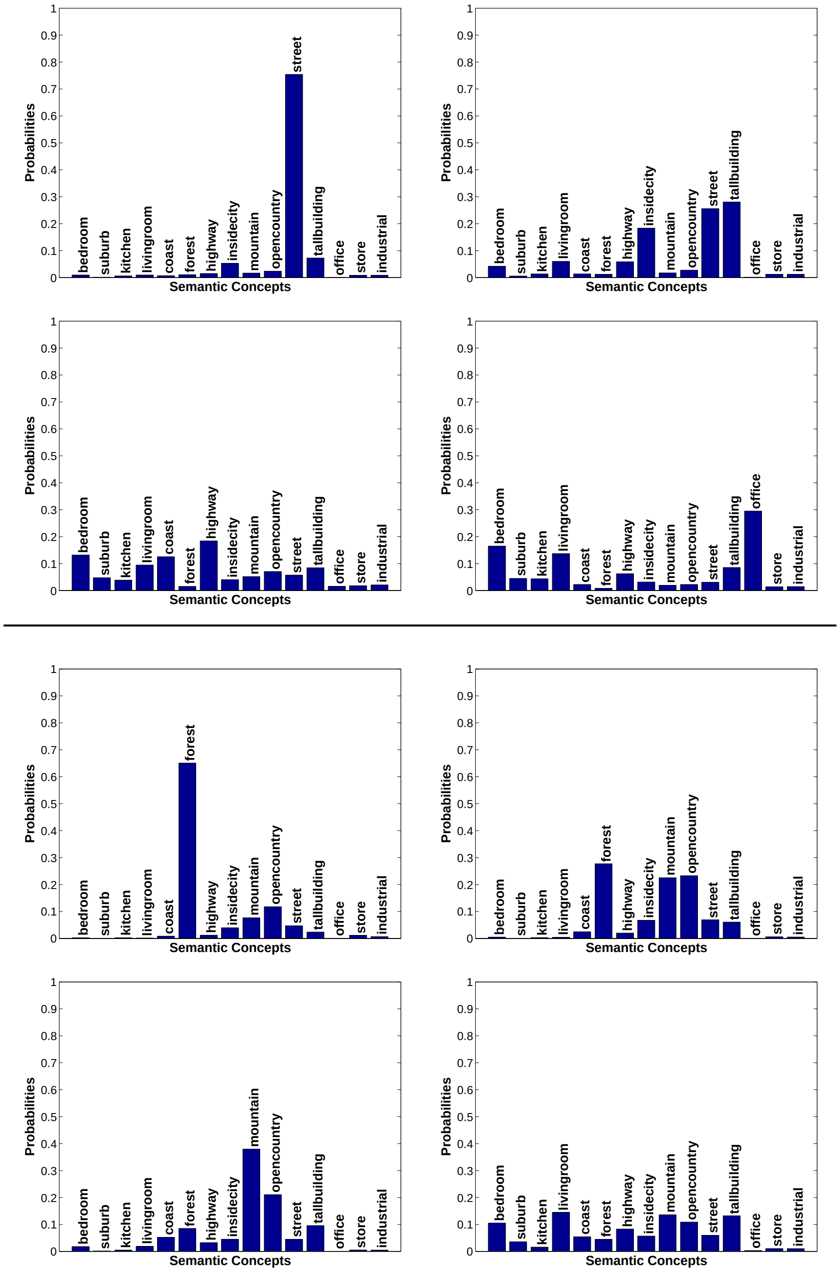


Figure 6.6: Four cluster centers for the class "street" (top) and "forest" (bottom). Note that each class comprises different co-occurrence patterns.

6.5.3 Choice of Appearance Features

6.2 compares the classification performance of the three appearance representations. In all cases, the contextual models yield improved performance, with a

Table 6.2: Impact of appearance space on classification accuracy.

Feature	Classification Accuracy (%)		Gain
	Appearance Models	Contextual Models	
SIFT-GRID using (2.21)	71.67 ± 1.17	**77.20 ± 0.39**	7.7%
SIFT-GRID using (2.23)	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%

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 6.5(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 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.

6.5.4 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 6.7 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 6.3.1, 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

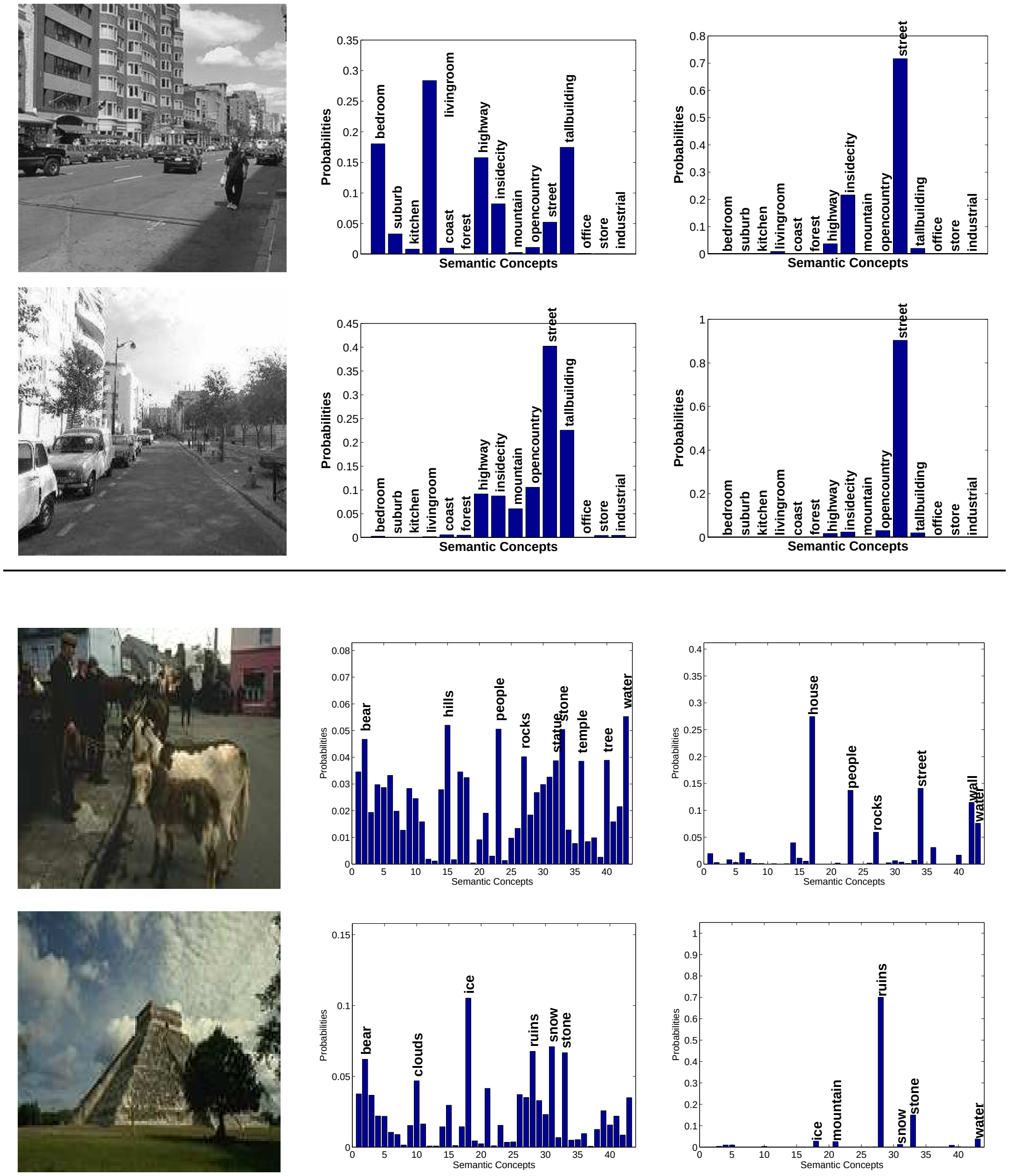


Figure 6.7: top) Two images from the "street" class of N15, and bottom) an image each from the "Ireland" and "Mayan ruins" CD of the Corel collection. Also shown with the images are the SMN and CMN vectors (middle and right column respectively). Notice that the CMN vectors are noise-free and capture the "gist" of the image.

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 image in the third row. 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”.

6.5.5 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 [159]. 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.

6.6 Comparison with Previous Work

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

Table 6.3: Classification Results on Natural Scene Categories.

Method	**Classif.**	**Dims.**[a]	**Accuracy (%)**
	N15 Dataset		
Contextual Models	Bayes	**15**	**77.20 ± 0.39**
pLSA [17][b]	SVM	40	72.7
pLSA [74]	SVM	60	63.3
LDA [77][e]	Bayesian	40	59.0
"gist" like [74]	SVM	16	45.3 ± 0.5
BoW [74]	SVM	400	74.8 ± 0.3
BoW [74]	SVM	200	72.2 ± 0.6
Bag of Concepts [83][c]	SVM	100	73.01
Kernel Codebook [154]	SVM	3200	~75[d]
Diffusion Distance [82]	SVM	2000	74.9
SIS [24]	SVM	200	74.94
Semantic Space [120]	SVM	15	73.95 ± 0.74

[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

Table 6.4: Classification Results on Natural Scene Categories.

Method	**Classif.**	**Dims.**[a]	**Accuracy (%)**
	N13 Dataset		
Contextual Models	Bayes	**13**	**80.86 ± 0.50**
LDA [77]	Bayesian	40	65.2
pLSA [17][b]	SVM	35	74.3
pLSA [114]	SVM	40	60.8
pLSA [74]	SVM	60	65.9
BoW [74]	SVM	200	74.7
Taxonomy [6]	Bayesian	40	68
“gist” features [65]	SVM	512	∼55[c]
Semantic Space [120]	SVM	13	77.57 ± 1.12

[a] Dimensionality of the space on which classification is performed

[b] Uses half of the dataset for training

[c] Accuracy estimated from figure

Table 6.5: Classification Results on Natural Scene Categories.

Method	**Classif.**	**Dims.**[a]	**Accuracy (%)**
	N8 Dataset		
Contextual Models	Bayes	**8**	**85.60 ± 0.70**
Context Ancestry [80]	Logistic	484	82
pLSA [17][b]	SVM	25	82.5
HDP-HMT [67]	Bayesian	200	84.5
“gist” [104][c]	SVM	512	83.7
Semantic Space [120]	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] Gist features implicitly uses weak spatial information

Table 6.6: 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 [74]	SVM	512	48.4
pLSA [17]	SVM	50	40.2
LDA [77]	Bayes	50	31.0
		C43 Dataset	
Contextual Models	Bayes	**43**	**42.9**
Appearance Models	Bayes	129	39.9
Bag of Words [74]	SVM	512	36.3
pLSA [17]	SVM	50	33.0
LDA [77]	Bayes	50	24.6

[a] Our implementation of the algorithms

6.6.1 Scene Classification

Given the posterior probabilities of (6.3), 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 k of largest posterior $P_{K|\Pi}(k|\pi)$. 6.3, 6.4 and 6.5 compare the resulting classification accuracies for N15, N13, and N8 respectively, 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 [77], pLSA [17, 114]), of which only [17] 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 BoW representation[1]. Note, however, that these require much higher dimensional spaces, e.g. a 400 visual-word vocabulary [74], 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[2]. Also reported is a baseline with discriminative learning [120] 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.

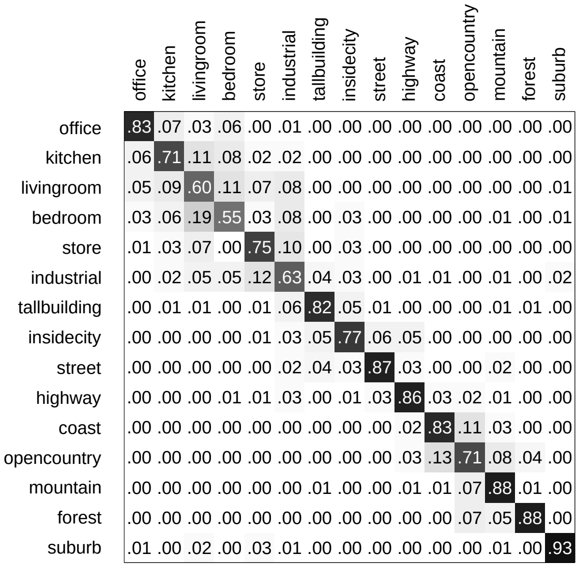
Within the area of context modeling, e.g. comparing to the methods of [104, 80], the proposed approach is again more effective. For the N8 (N13, N15) dataset, [104] ([65], [74]) report a classification accuracy of 83.7% (55%, 45.3%[3]), respectively, using the "gist" features of [104]. The corresponding figures for the proposed contextual models are 85.6% (80.86%, 77.2%). The scene confusion matrix for N15 is also shown in Figure 6.8. Note that most errors are due to confusion between "coast" and "open country," "living room" and "bed room," or "living room" and "kitchen." These are very tolerable errors, given the similarity of scenes in these classes. In fact, their images are sometimes difficult to discriminate even for a human.

Finally, 6.6 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 [74, 17] the proposed contextual modeling again achieves significantly

[1] Note that BoW representation is obtained by vector quantizing the space of descriptors and representing an image with a visual word histogram.

[2] We note that better results have been reported for an extension of the BoW representation that includes a weak encoding of spatial information [74, 179]. These results are the current state-of-the-art for N15: 81.4% [74] using a SVM classifier on an 8400 dimensional space; 85.2% [179] 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.

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



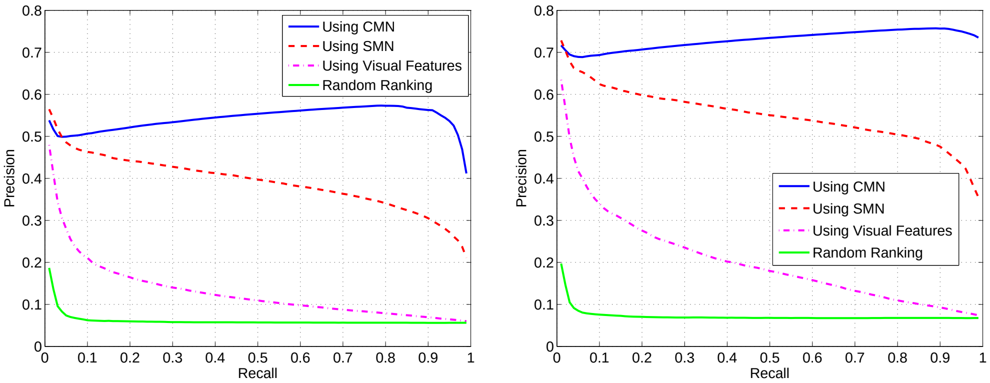
	office	kitchen	livingroom	bedroom	store	industrial	tallbuilding	insidecity	street	highway	coast	opencountry	mountain	forest	suburb
office	.83	.07	.03	.06	.00	.01	.00	.00	.00	.00	.00	.00	.00	.00	.00
kitchen	.06	.71	.11	.08	.02	.02	.00	.00	.00	.00	.00	.00	.00	.00	.00
livingroom	.05	.09	.60	.11	.07	.08	.00	.00	.00	.00	.00	.00	.00	.00	.01
bedroom	.03	.06	.19	.55	.03	.08	.00	.03	.00	.00	.00	.00	.01	.00	.01
store	.01	.03	.07	.00	.75	.10	.00	.03	.00	.00	.00	.00	.00	.00	.00
industrial	.00	.02	.05	.05	.12	.63	.04	.03	.00	.01	.01	.00	.01	.00	.02
tallbuilding	.00	.01	.01	.00	.01	.06	.82	.05	.01	.00	.00	.00	.01	.01	.00
insidecity	.00	.00	.00	.00	.01	.03	.05	.77	.06	.05	.00	.00	.00	.00	.00
street	.00	.00	.00	.00	.00	.02	.04	.03	.87	.03	.00	.00	.02	.00	.00
highway	.00	.00	.00	.01	.01	.03	.00	.01	.03	.86	.03	.02	.01	.00	.00
coast	.00	.00	.00	.00	.00	.00	.00	.00	.00	.02	.83	.11	.03	.00	.00
opencountry	.00	.00	.00	.00	.00	.00	.00	.00	.00	.03	.13	.71	.08	.04	.00
mountain	.00	.00	.00	.00	.00	.00	.01	.00	.00	.01	.01	.07	.88	.01	.00
forest	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.07	.05	.88	.00
suburb	.01	.00	.02	.00	.03	.01	.00	.00	.00	.00	.00	.00	.01	.00	.93

Figure 6.8: Class confusion matrix for classification on the N15 dataset. The average accuracy is 77.20%

higher accuracy. On C50, its accuracy is 57.8% while [74] and [17] 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.

6.6.2 Image Retrieval Performance

Finally, the benefits of holistic context modeling were evaluated on the task of content based image retrieval, using the 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 Chapter 3, 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 [119].



(a) F18

(b) C15

Figure 6.9: Precision-recall curves achieved with SMN, CMN, visual matching and chance level image retrieval.

Figure 6.9, presents precision-recall (PR) curves on C15 and F18. Also shown are the performance of the image matching system of [156], 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 6.10 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.

6.7 Acknowledgments

The text of Chapter 6, in full, is based on the material as it appears in: N. Rasiwasia and N. Vasconcelos, *'Holistic Context Models for Visual Recognition'*, Accepted to appear in IEEE Transactions on Pattern Analysis and Machine Intelligence, N. Rasiwasia and N. Vasconcelos, *'Holistic Context Modeling using Semantic Co-occurrences'*, IEEE Conference on Computer Vision and Pattern Recognition, Miami, June 2009, and N. Rasiwasia and N. Vasconcelos, *'Image Retrieval using Query by Contextual Example'*, ACM Conference on Multimedia Information Retrieval, pp. 164-171, Vancouver, Oct 2008. The dissertation author was a primary researcher and an author of the cited material.

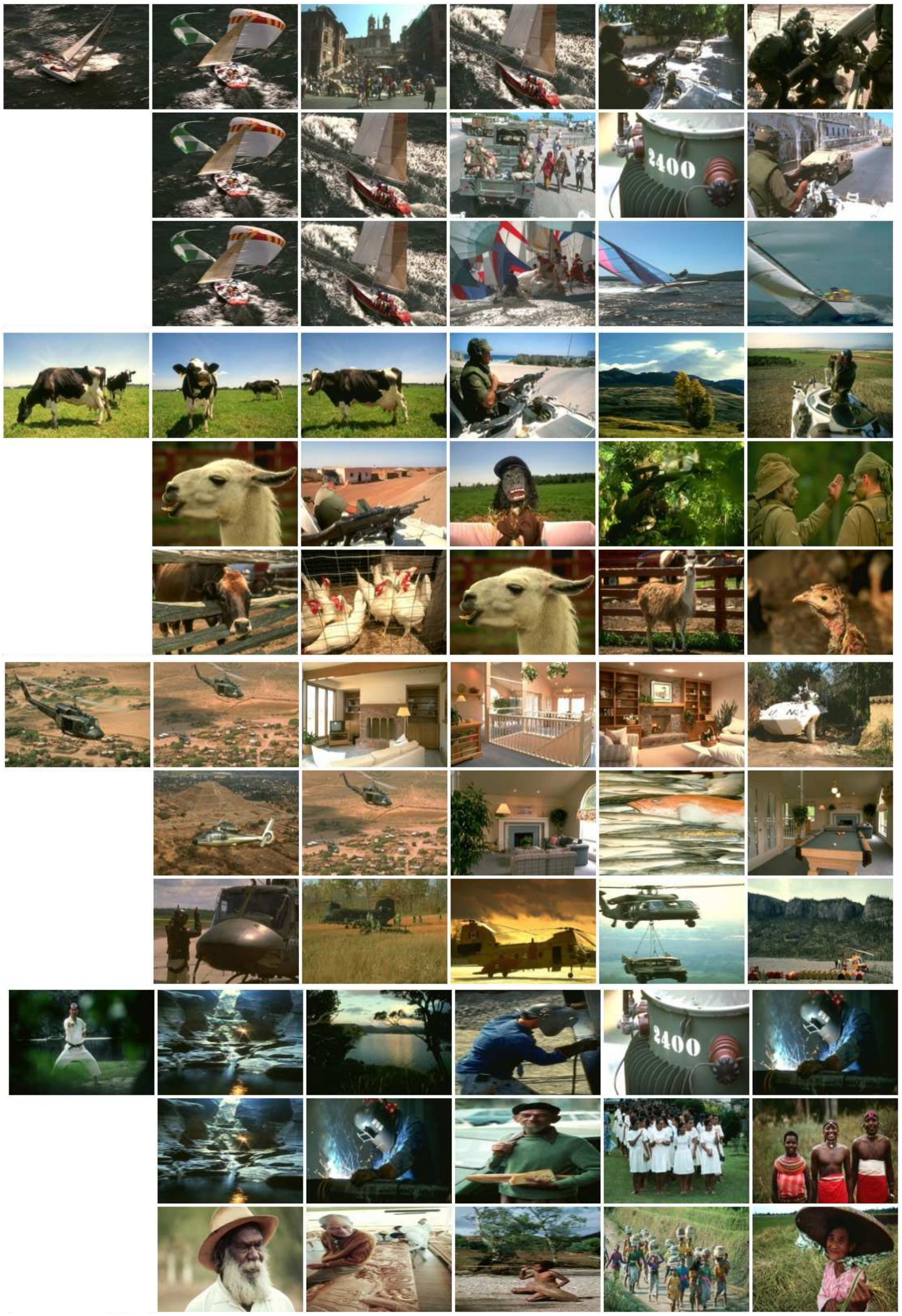


Figure 6.10: Retrieval results for four image queries shown on the left-most column. The first, second, and third row of every query show the five top matches using image matching, SMN, and CMN-based retrieval, respectively.