Chapter 4

Scene Classification with Semantic Representation

In this chapter we introduce the problem of scene classification and present a novel solution based on semantic image representation.

4.1 Introduction

Scene classification is an important problem for computer vision, and has received considerable attention in the recent past. It differs from the conventional object detection/classification, to the extent that a scene is composed of several entities often organized in an unpredictable layout[113]. Images of scenes also differ from images of objects with respect to the distance between the camera and the elements in the image [104]. For a given scene, it is virtually impossible to define a set of properties that would be inclusive of all its possible visual manifestations. Frequently, images from two different scene categories are visually similar, e.g., it can be difficult to distinguish between scenes such as "open country" and "mountain" (see Sec. 4.4).

Early efforts at scene classification targeted binary problems, such as distinguishing indoor from outdoor scenes [142], city views from landscape etc. Subsequent research was inspired by the literature on human perception. In [9], it was shown that humans can recognize scenes by considering them in a "holistic" manner, without recognizing individual objects. Recently, it was also found that humans can perform high-level categorization tasks extremely rapidly [144] in the near absence of attention [78]. Drawing inspiration from the perceptual literature, [104] proposes a low dimensional representation of scenes, based on several global properties such as "naturalness", "openness", etc. More recently, there has been an effort to solve the problem in greater generality, through design of techniques capable of classifying relatively large number of scene categories [166, 77, 113, 74, 16, 83], and a dataset of 15 categories has been used to compare the performance of various systems[74, 83]. These methods tend to rely on *local region descriptors*, modeling an image as a bag-of-features (BoF, see Section 2.1.1. The space of local region descriptors is then quantized, based on some clustering mechanism, and the mean vectors of these clusters, commonly known

as "visual-words"[1] are chosen as their representatives, thereby yielding the bag-of-words (BoW) representation. This representation is motivated by the time-tested BoW model, widely used in text-retrieval [125]. The analogy between visual-words and text-words is also explored in [130].

Lately, various extensions of this basic BoW model have been proposed [77, 113, 16, 83]. All such methods aim to provide a compact lower dimensional representation using some intermediate characterization on a latent space, commonly known as the intermediate "theme" or "topic" representation [77]. The rationale is that images which share frequently co-occurring visual-words have similar representation in the latent space, even if they have no visual-words in common. This leads to representations robust to the problems of polysemy - a single visual-words may represent different scene content, and synonymy - different visual-words may represent the same content [113]. It also helps to remove the redundancy that may be present in the basic BoW model, and provides a semantically more meaningful image representation. Moreover, a lower dimensional latent space speeds up computation: for example, the time complexity of a Support Vector Machine (SVM) is linear in the dimension of the feature space. Finally, it is unclear that the success of the basic BoW model would scale to very large problems, containing both large image corpuses and a large number of scene categories. In fact, this has been shown not to be the case in text-retrieval, where it is now well established that a flat representation is insufficient for large scale systems, and the use of intermediate latent spaces leads to more robust solutions [58, 14]. However, a direct translation of these methods to computer vision has always incurred a loss in performance, and latent models have not yet been shown to be competitive with the flat BoW representation [83, 74].

In this chapter we propose an alternative solution, based on semantic image representation. Like the latent model approaches we introduce an intermediate space - the semantic space, however, instead of learning the themes in an unsupervised manner from the BoW representation as is done in existing works, the

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

semantic themes are explicitly defined and the images are casually annotated with respect to their presence. This can *always* be done since, in the absence of "thematic" annotations, the "themes" can be made equal to the class labels, which are always available. The number of semantic themes used defines the dimensionality of the intermediate theme space, henceforth referred to as "semantic space". Each theme induces a probability density on the space of low-level features, and the image is represented as the vector of posterior theme probabilities. An implementation of this approach is presented and compared to existing algorithms on benchmark datasets. It is shown that the proposed low dimensional representation outperforms the unsupervised latent-space approaches, and achieves performance close to the state of the art, previously only accessible with the flat BoW representation using a much higher dimensional image representation.

The paper is organized as follows. Section 4.2 discusses related work. Section 4.3 presents the approach now proposed, and Section 4.4 an empirical evaluation on benchmark datasets, allowing comparison to previous results.

4.2 Related Work

Low dimensional representations for scene classification have been studied in [77, 113, 16, 83]. On one hand, it is noticed that increasing the size of the codebook improves classification performance[102]. Csurka et al. [27] compare different codebook sizes ranging from 100 to 2500 visual-words, showing that performance degrades monotonically as size decreases. They choose a size of 1000, based on a trade-off between accuracy and speed. Quelhas et al. [113] also experience a monotonic degradation of performance for 3-class classification, and use a codebook of 1000 visual-words. In [74], Lazebnik et al. show that performance increases when codebook size is increased from 200 to 400 visual-words.

On the other hand, there is a strong desire for low dimensional representations, for the benefits elucidated in Section 4.1. This is achieved by resorting to techniques from the text-processing literature, such as Latent Dirichlet Allocation (LDA) [14], Probabilistic Latent Semantic Analysis (pLSA) [58] etc., which

produce an intermediate latent "theme" representation. Fei-Fei et al. [77] motivate the use of intermediate representations, citing the use of "textons" in texture retrieval. They then propose two variations of LDA to generate the intermediate theme representation. In [113], Quelhas et al. use pLSA, to generate the compact representation. They argue that pLSA has the dual ability to generate a robust, low dimensional scene representation, and to automatically capture meaningful scene aspects or themes. pLSA is also used by Bosch et al. in [16]. Another approach to two-level representation based on the Maximization of Mutual Information (MMI) is presented in [83]. However, a steep drop in classification performance is often experienced as a result of dimensionality reduction [83, 74].

4.3 Proposed Approach

A scene classification system can be broadly divided into two modules. The first defines the image representation, while the second delineates the classifier used for decision making. Since the main goal of this thesis is to present a low-dimensional semantic theme representation, we do not duel on the choice of classifier, simply using an SVM. This is the standard choice in the scene classification literature [166, 102, 27].

Semantic Theme Representation

Under the proposed classification framework, an image is represented by its semantic multinomial (SMN). This is similar in principle to the two level image representations of [77, 113, 16], where an intermediate "theme" space is learned in an unsupervised fashion. In the proposed formulation the semantic space serves as the surrogate for the intermediate "theme" space. As discussed in Chapter 2, learning a semantic space requires a vocabulary of semantic concepts $\mathcal{L}$ and a dataset annotated with respect to $\mathcal{L}$. These semantic concepts serve the same role as the intermediate "themes" in the existing work[77, 113, 16]. In general, semantic concepts or "themes" are different from image classes. For example, images in the "Street" class of Figure 4.2i contain themes such as "Road", "Sky", "People", or

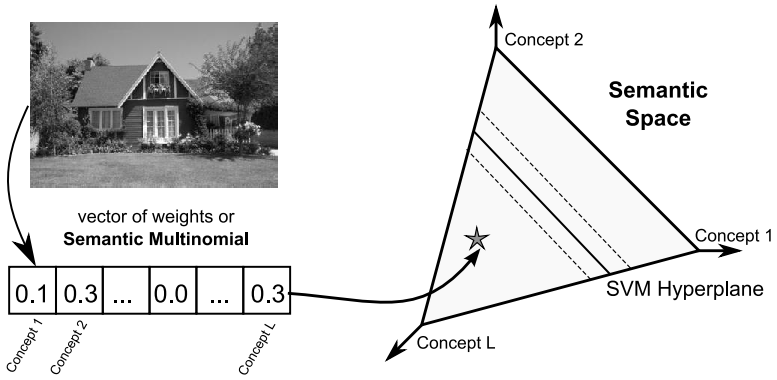


Figure 4.1: The proposed scene classification architecture.

"Cars". However, current popular scene classification datasets lack such semantic theme annotations and in the absence of these, the set of scene categories $\mathcal{W} = \{1, \ldots, K\}$, e.g. "Street", can serve as a proxy for $\mathcal{L}$. In this case, each image is only explicitly annotated with one "theme", even though it may depict multiple: e.g. most images in the "Street" class of Figure 4.2i also depict "Buildings". We refer to this limited type of scene labeling as *casual annotation*. This is the annotation mode for all results reported in this paper, to enable comparison to previous scene classification work. We will see that supervised learning of the intermediate theme space with casual annotations can be far superior to unsupervised learning of a latent theme space, as previously proposed [77].

Scene Classification

Due to the limited information contained in casual annotations, images cannot be simply represented by the caption vectors $\mathbf{c}_i$. In fact, $\mathbf{c}_i$ is only available for training images, and $\mathbf{c}_{i,j} = 0$ does not mean that the i^{th} image does not contain the j^{th} theme, simply that it was not annotated with it. Instead, the proposed classification system represents images by vectors of theme frequency, or counts. In this way, an image can be associated with multiple themes, even when there are no multiple associations in the labels used for training. As shown in Figure 4.1, the scene classifier (e.g. SVM) then operates on this feature space.

4.4 Experimental evaluation

We now present an empirical evaluation of the model as a low dimensional semantic theme representation for two publicly available datasets, comparing performance with [83, 16, 77, 74]. We also present a study of classification accuracy as a function of semantic space dimensions.

4.4.1 Datasets

Scene classification results are presented on two public datasets: 1) Natural15 [74] and 2) Corel50 photos, used in [21] for image annotation comprising of 50 scene categories. The details of these datasets are discussed in Appendix A.1.1 and Appendix A.1.3 respectively. The use of the Natural15 dataset allow us to directly compare with the existing results on scene classification. In particular, we show a comparison of our results using low-dimensional representation with those of [83, 74, 77, 16]. The Corel50 dataset has 100 high resolution images per category, which we resize to an average of 180×120 pixels. To the best of our knowledge, this is the database with maximum number of scene categories so far studied in the literature (viz. 50). Since the dimension of our semantic theme representation directly depends on the number of scene categories (see Sec. 4.3), this dataset enables the study of the effects of dimensionality as the number of categories grows.

4.4.2 Experimental Protocol

At the low level, images are represented as bags of 8×8 vectors of discrete cosine transform (DCT) coefficients sampled on a uniform grid. The Corel50 dataset consists of color images which are converted from RGB to YcrCb colorspace [2]. The Natural15, consist of grayscale images hence no such conversion is required. Semantic theme densities are learned on a 36(out of 64) / 64(out of 192) dimensional subspace of the DCT coefficients for Natural15 and Corel50 dataset respectively,

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

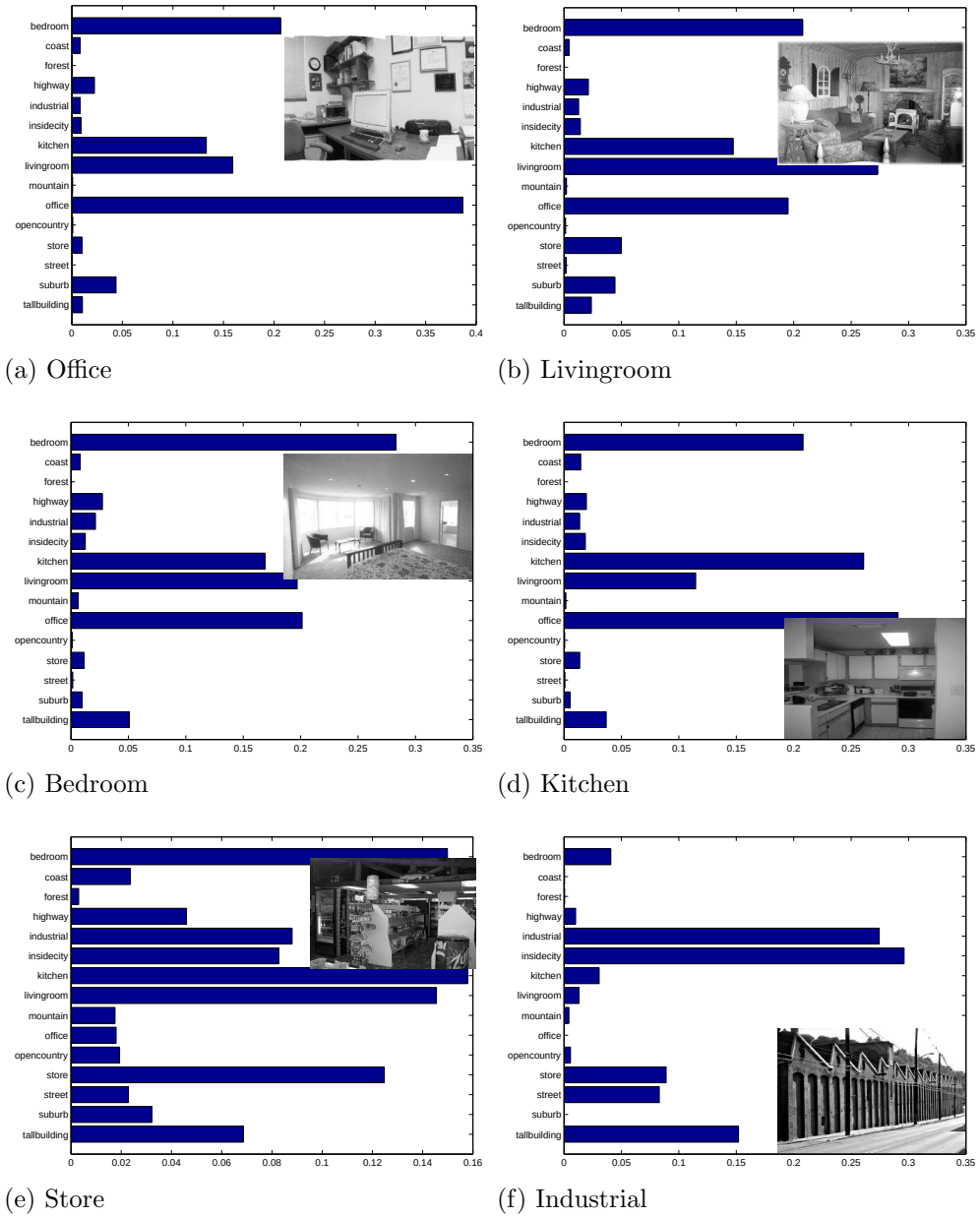
with each theme modeled as a mixture of 128 Gaussian components. The images at the semantic theme level are represented by 15 (50) dimensional theme vectors for Natural15 (Corel50). Later on, we also show that not all 50 themes are equally informative on Corel50. 100 (90) images per scene are used to learn the theme density for Natural15 (Corel50), and the rest of the images are used as the test set. All experiments on Natural15 are repeated 5 times with different randomly selected train and test images. For Corel50 dataset, we use the same training and test images as used in [21, 35]. A multi-class SVM using one-vs-all strategy with Gaussian kernel is used for classification, with the parameters obtained by 3-fold cross validation.

4.4.3 Results

We start by studying scene classification accuracy.

Scene classification

Figure 4.2 shows an example from each of the fifteen scene categories of Natural15, along with their semantic theme representation. All images shown are actually classified correctly by the classifier. Two interesting observations can be made: 1) semantic theme vectors *do capture* the different semantic meanings of the images, hence correlating well with human perception. For example, the theme vector shown for the scene from the category "Forest" in Figure 4.2n, has large weights for themes such as "*forest*", "*mountain*" and "*open-country*", which are suitable themes for the scene, and 2) in many examples (viz. Figure 4.2(d)-(f),(h),(i)), even though the semantic theme corresponding to the same semantic scene category does not have the highest probability, the scene is still classified correctly. For example in Figure 4.2i, in spite of the "*street*" theme having much lower probability than "*tall-building*", "*inside-city*", "*highway*", the image is classified as belonging to the "Street" category. This is a direct consequence of the classifier learning *associations* between themes, despite the casual nature of the annotations. Figure 4.4 presents some of the misclassified images from the worst performing scene categories, along with the scene category they are classified into.

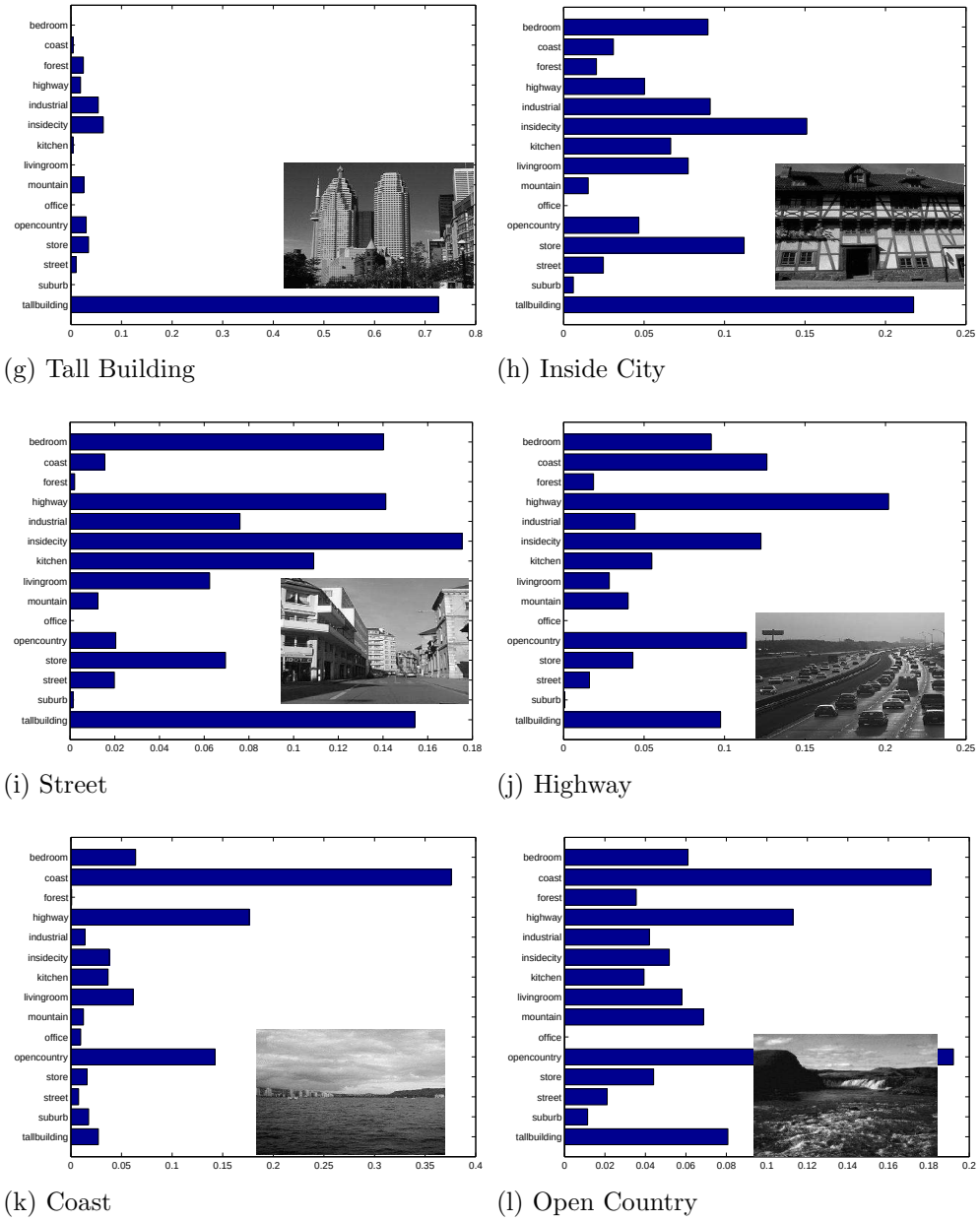


(a) Office (b) Livingroom

(c) Bedroom (d) Kitchen

(e) Store (f) Industrial

Figure 4.2: Theme vectors from each of the scenes of fifteen scene categories.

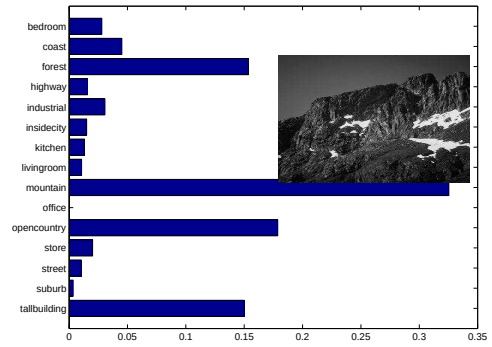


(g) Tall Building (h) Inside City

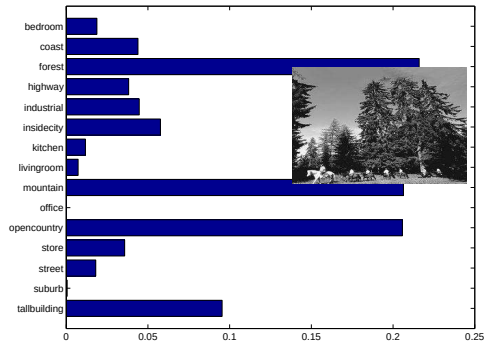
(i) Street (j) Highway

(k) Coast (l) Open Country

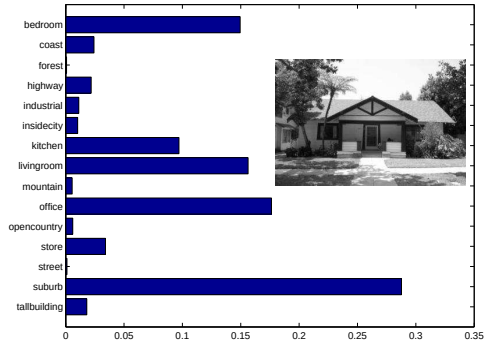
Figure 4.2: Theme vectors from each of the scenes of fifteen scene categories. (continued)



(m) Mountain



(n) Forest



(o) Suburb

Figure 4.2: Theme vectors from each of the scenes of fifteen scene categories. (continued)

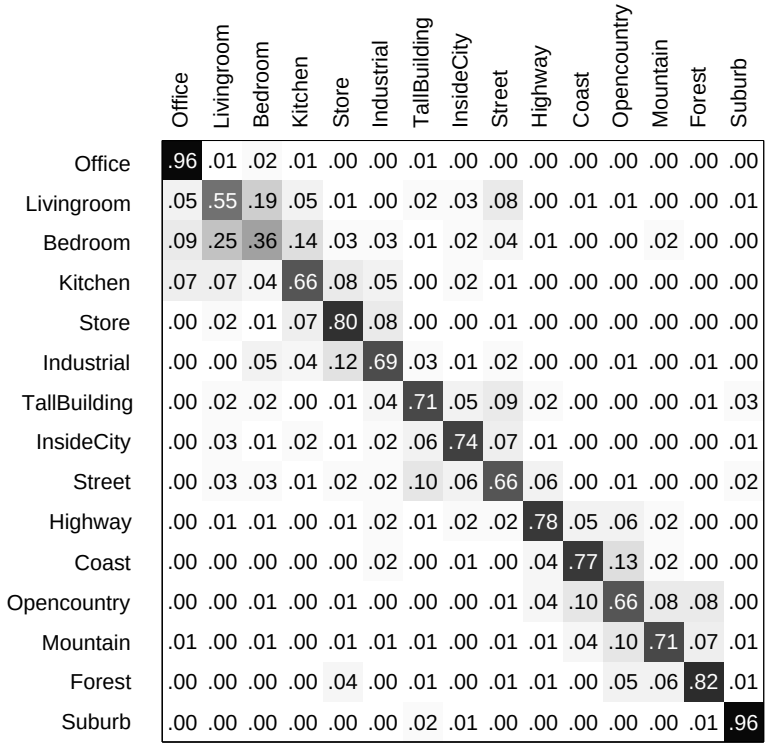


Figure 4.3: Confusion Table for our method using 100 training image and rest as test examples from each category of Natural15. The average performance is 72.2% $\pm$ 0.2

The confusion table for Natural15 is shown in Figure 4.3. The average classification accuracy, over all categories is $72.2 \pm 0.2\%$. As was experienced by [74], there is confusion between indoor categories such as "Bedroom", "Livingroom" and "Kitchen" and outdoor categories like "Opencountry" and "Mountain". In fact close to 25% of images from the category "Bedroom" were classified as "Livingroom". On Corel50, the classification accuracy stands at 56.8%, the chance classification accuracy being 2%. Figure 4.5 shows some of the images from various scene categories of Corel50 dataset. Also shown in Figure 4.6 is the theme vector for the image of Figure 4.5(a).

Comparison with existing work

4.1 compares the classification accuracy of the proposed method on Natural15, using 15 dimensional theme vectors, with the existing results in the literature. It is evident that when compared to the MMI based dimensionality reduction of Liu et at. [83], which achieves a rate of 63.32% using a 20 dimensional space, the method performs substantially better, achieving a rate of 72.2% on an even lower dimensional space of 15 themes. The performance is equal to that of Lazebnik et al. [74][3], who represent images as the basic BoW model, using 200 visual-words. A similar comparison on the thirteen subcategories of the dataset used in [77, 16], is presented in 4.1. Again, the proposed low-dimensional theme vector based representation performs close to the best results in the literature, with a much lower dimensional space. This dataset also shows that the proposed method substantially outperforms the latent-space method of Fei-Fei et al. [77], and achieves equivalent performance the latent-space method of Bosch et al. [16] with roughly half of its dimensionality.

Informative semantic themes

In all the experiments conducted above, scene categories served as a proxy for the intermediate themes. This is a practical approach to scene classification where the images are devoid of other annotations. However, it might seem that the extension of the current framework to very large-scale problems involving thousands of categories, will annul the benefits gained by the proposed representation, as the dimension of the semantic space would grow with the number of categories. The effects of varying the dimensions of the semantic space on the classification accuracy is studied, on Corel50 dataset. Semantic spaces of k dimensions were produced by ordering the semantic themes by the variance of their posterior probabilities, and selecting the k of largest variance (for k ranging from 2 to 50). Clas-

[3]Note that the best results on this dataset, are obtained by incorporating spatial information, and representing images as histograms at different spatial resolution, with Spatial Pyramid Matching [74]. The accuracy is 81.1%, with a 4200 dimensional feature space. Multi-resolution semantic representations would also be possible with the proposed method, as well as the incorporation of spatial information, but these extensions are beyond the scope of the current discussion.

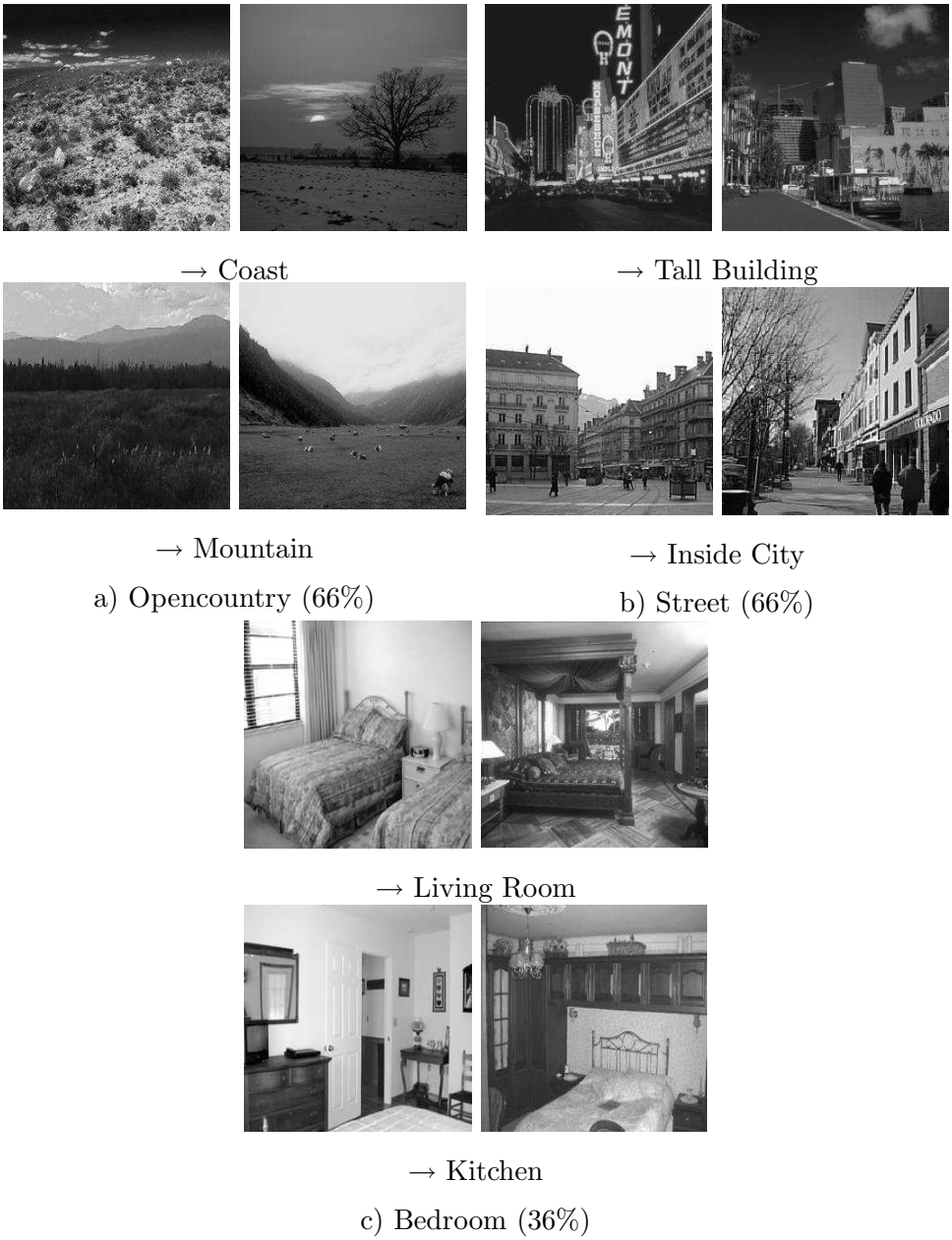


Figure 4.4: Some images from worst performing scene categories in Natural15. (→) implies the scene category the image is classified into.

sification was performed on each of these resulting spaces and Figure 4.7 presents the performance as a function of the dimension. It can be observed that not all of the 50 dimensions are equally informative, as moving from 40 to 50 dimensions increases performance by only 3.8% (a relative gain of 6.7%). This can be explained

→ California Coast
(a) California Coast

→ Trains
(b) Trains

→ Flowers
(c) Peru

→ Rural France
(d) Africa

→ Tropical Plants
(e) Birds

→ Auto Racing
(f) South Africa

Figure 4.5: Some images from the Corel50 dataset. (→) implies the scene category the image is classified into. (a) and (b) show two examples of correctly classified images, (c) and (d) two reasonably misclassified images and (e) and (f) shows two examples of error.

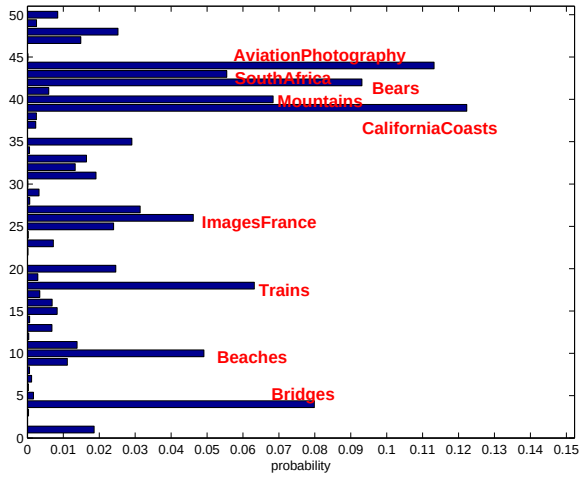


Figure 4.6: The theme vector for the image in Figure 4.5(a).

by the plot of variance of the posterior probabilities for the 50 themes (in the same figure). For very large scale problems, where most of the variance is expected to be captured by a subset of the features, the correlation of classification performance with the variance of the themes indicates that the number of informative themes

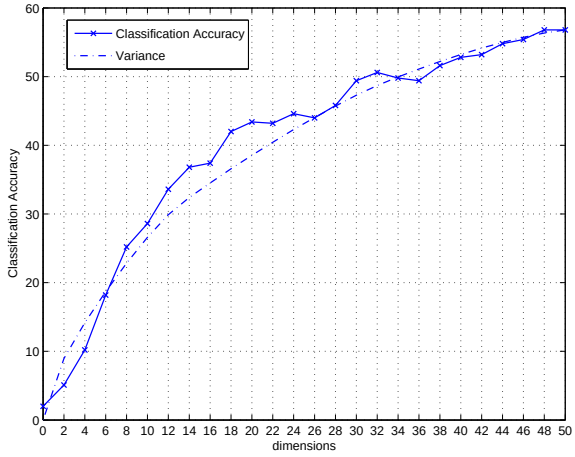


Figure 4.7: Classification performance as a function of the semantic space dimensions. Also shown, is the growth of the variance of the semantic themes, scaled appropriately.

Table 4.1: Classification Result for 15 scene categories.

Method	**Dimensions**	**Classification Accuracy**
Our method	**15**	**72.2 ± 0.2**
Liu et al. [83]	**20**	**63.32**
Liu et al. [83]	200	75.16
Lazebnik et al. [74]	200	72.2 ± 0.6

would grow sub-linearly as the number of scene categories is increased. It is unclear that this type of behavior will hold for the flat BoW representations. In the works previously presented in the literature, the codebook has *linear* size on the number of classes.

The results presented above allow a number of conclusions. While low dimensional semantic representations are desirable for the reasons discussed in Section 4.1, previous approaches based on latent-space models have failed to match the performance of the flat BoW model, which has high dimensionality. We have shown that this is indeed possible, with methods that have much lower complexity than the latent-space approaches previously proposed, but make better use of

Table 4.2: Classification Result for 13 scene category subset.

Method	**Dimensions**	**Classification Accuracy**
Our method	**13**	**72.7 ± 0.3**
Bosch et al. [16]	25	73.4
Fei-Fei et al. [77]	40	65.2
Lazebnik et al. [74]	200	74.7

the available labeling information. We have also shown that the proposed method extracts meaningful semantic image descriptors, despite the casual nature of the training annotations, and is able to learn co-occurrences of semantic themes without explicit training for these. Finally a study of the effect of dimensionality on the classification performance was presented, and indicated that the dimensionality would grow sub-linearly with the number of scene categories. This could be a significant advantage over the flat BoW model which, although successful for the limited datasets in current use, will likely not scale well when the class vocabulary increases.

4.5 Acknowledgments

The text of Chapter 4, in part, is based on the material as it appears in: Rasiwasia, N., Vasconcelos, N. "Scene Classification with Low-dimensional Semantic Spaces and Weak Supervision" IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Anchorage, June 2008 The dissertation author was a primary researcher and an author of the cited material.