Mining Discriminative Adjectives and Prepositions for Natural Scene Recognition

Bangpeng Yao[1], Juan Carlos Niebles[2,3], Li Fei-Fei[1]

[1]Department of Computer Science, Princeton University, NJ 08540, USA
[2]Department of Electrical Engineering, Princeton University, NJ 08544, USA
[3]Robotics and Intelligent Systems Group, Universidad del Norte, Colombia

{bangpeng,feifeili}@cs.princeton.edu jniebles@princeton.edu

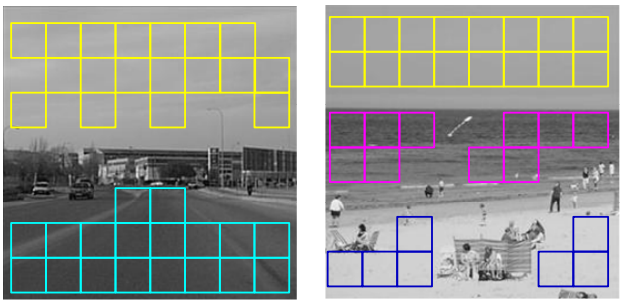
Abstract

This paper presents a method that considers not only patch appearances, but also patch relationships in the form of adjectives *and* prepositions *for natural scene recognition. Most of the existing scene categorization approaches only use patch appearances or co-occurrence of patch appearances to determine the scene categories, but the relationships among patches remain ignored. Those relationships are, however, critical for recognition and understanding. For example, a 'beach' scene can be characterized by a 'sky' region* above *'sand', and a 'water' region* between *'sky' and 'sand'. We believe that exploiting such relations between image regions can improve scene recognition. In our approach, each image is represented as a spatial pyramid, from which we obtain a collection of patch appearances with spatial layout information. We apply a feature mining approach to get discriminative patch combinations. The mined patch combinations can be interpreted as* adjectives *or* prepositions*, which are used for scene understanding and recognition. Experimental results on a fifteen class scene dataset show that our approach achieves competitive state-of-the-art recognition accuracy, while providing a rich description of the scene classes in terms of the mined adjectives and prepositions.*

1. Introduction

In this paper, we address the problem of natural scene classification and understanding. We aim to develop algorithms that can recognize scene classes such as highway, mountain, or living room from input images, as well as learn the types of image patches and the relationships among such patches that define each scene category.

A common strategy for scene classification is based on the bag-of-words image representation [17, 2]. However, although this approach has demonstrated competitive performance, its descriptive and discriminative abilities are severely limited, partly because the bag-of-words representation ignores the geometrical arrangement of patches in the image.



(a) Highway (b) Coast

Figure 1. Examples showing the discriminative and descriptive ability of our approach. Both the "highway" and "coast" images have sky regions on the top (yellow patches). Using prepositions, the highway image can be described as having a region of sky *above* a region of road (cyan patches), while the coast image can be described as having a region of sky *above* sand (blue patches) and water (magenta patches) in between.

One promising strategy to obtain more discriminative features for scene classification and understanding, consists of exploring the spatial relationships among image regions. For example, in Figure 1(b), a 'coast' image can be described as having a 'sky' region *above* a 'sand' region, and a 'water' region *between* 'sky' and 'sand'. This representation is capable of not only discriminating between 'coast' and 'highway' (Figure 1(a)) images, but also understanding the intrinsic properties of 'coast' images. In this work, we aim to explore such relationships for scene classification and understanding. We consider relationships of two

978-1-4244-3993-5/09/$25.00 ©2009 IEEE

types: *adjectives* (e.g. smaller, brighter) and *prepositions* (e.g. above, left).

In general, modeling adjectives and prepositions for images can be computationally challenging due to the large number of possible adjectives and prepositions. In our approach, we adopt an image representation inspired by the spatial pyramid method [8]. We partition the image into increasingly finer subregions, and compute a histogram of local features on each subregion, similarly to [8]. Furthermore, instead of considering local features of a fixed size as in [8], we use local image patches at multiple scales. Our method can represent multiple relationships among local patches which capture their appearances as well as spatial layout. Based on this multi-scale image representation with spatial layouts, we apply a data mining approach [1] to obtain all discriminative patch relationships for each class. Those relationships can be interpreted as adjectives and prepositions. The data mining approach is suitable in our problem due to its ability to deal with large amounts of data efficiently.

The remaining part of this paper is organized as follows. Section 2 overviews related work. Our image representation and feature mining approaches are described in Section 3 and Section 4 respectively. The classification approach used in our method is presented in Section 5. Finally, we present experimental results on a natural scene data set in Section 6.

2. Previous Work

In order to encode spatial information in images, several methods have considered using feature compositions to classify images (e.g. 'doublets' [15] and 'correlatons' [14]). The discriminative ability of these approaches is limited due to their reliance on feature co-occurrences only. Several other approaches based on generative part-based models (e.g. Pictorial Structures [3] and Constellation models [4, 18]) usually entail significant computational cost, and are therefore limited to a small number of image parts (typically less than 10).

Recently, researchers have proposed the application of graphical models to the problem of learning spatial relationships among image regions [13, 11]. These approaches first use segmentation algorithms to obtain image regions, and then learn the relationships among image regions via graphical modeling. In the graphical models, each node represents one image patch, and each edge represents one relationship between two patches. This line of approaches can perform object/scene recognition and localization simultaneously and achieve promising performance on many data sets. However, they rely on image segmentation as a preprocessing step. Although it has been shown that current segmentation can already facilitate image recognition [10], obtaining semantically meaningful image segmentations is still an active research topic in computer vision.

Instead of using generative models and relying on image segmentation as pre-processing stage, in this paper we aim to discover discriminative relationships to understand scene classes. One work closely related to ours is the spatial pyramid matching method [8]. In their framework, the image is partitioned into increasingly finer sub-regions and histograms of local features are computed inside each subregion. The similarity between two histograms is measured by the pyramid match kernel [5]. This approach shows promising performance in scene and object recognition tasks, partly due to its ability to encode weak spatial relationships among image patches and the effectiveness of the pyramid match kernel. However, the spatial pyramid representation does not explicitly capture stronger relationships (such as adjectives and prepositions) among patches in the image.

Another work for adjectives and prepositions modeling is the method of Gupta and Davis [6]. On a set of weakly labeled images, this approach uses relationship words (adjectives and prepositions) to reduce correspondence ambiguity and applies a constrained model to achieve better image labeling performance. But the adjectives and prepositions considered in this approach are manually pre-defined rather than automatically discovered. In this paper, the adjectives and prepositions can be automatically obtained given a set of training images.

Our method is based on a data mining approach [1]. Due to its computational efficiency, data mining has been recently used to mine feature configurations for object and action recognition [12, 19]. Our method differs from these approaches in that our mining procedure considers more precise spatial information.

3. Adjectives and Prepositions in Images

In this section, we first introduce our image representation which is based on a spatial pyramid and local patches of multiple scales. We then define adjectives and prepositions within the scope of our framework. We also show that our approach can cover a large volume of patch relationships, which instantiate the adjectives and prepositions under consideration.

3.1. Multi-Scale Local Patches and Spatial Pyramid Image Representation

The goal of feature representation is to obtain an image description which is robust to within-class variations and can discriminate images from different classes. In addition, the feature representation should enable us to extract adjectives and prepositions effectively. We achieve the goals by representing images in a multi-scale (or 'coarse-to-fine') manner [16] and adopting a modified spatial pyramid technique [8]. Figure 2 provides an illustration of our represen-

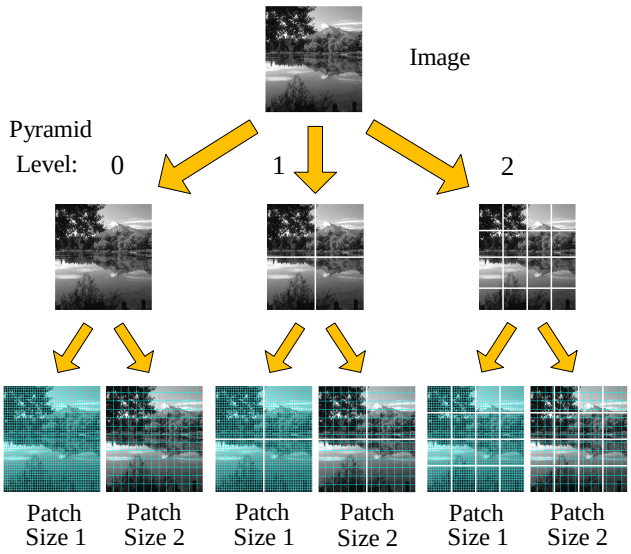


Figure 2. Image representation based on multi-scale local patches and a spatial pyramid. This figure shows an example that uses a 3-level spatial pyramid and local patches with 2 different scales.

tation scheme.

Our image representation is constructed as follows. We adopt a modified spatial pyramid representation [8], which captures the spatial layout of patches in the image. A spatial pyramid is constructed by sequentially dividing an image into increasingly finer sub-regions. Figure 2 shows an example of a 3-level spatial pyramid. In the l-th level ($l \in \{0, 1, 2\}$ in Figure 2) of the spatial pyramid, the image is divided into $2^l \times 2^l$ subregions. We extract local image patches of multiple scales from each subregion, which are obtained by densely sampling the image using a regular grid. In our example of Figure 2, we use two different local patch sizes. In order to represent the appearance of each patch, we first extract SIFT descriptors [9] and then assign each patch to one codeword. The visual dictionary of codewords can be obtained by clustering patch descriptors from the set of training images. Finally, the image is represented by a set of histograms[1], where each histogram counts the occurrence of visual codewords of a particular scale over one subregion in the spatial pyramid.

Implementation Details. In our experiments, we use a three-level spatial pyramid, i.e., we partition the image into 1×1, 2×2, and 4×4 sub-regions respectively. Overlapping image patches of size 8×8, 16×16, and 24×24 are extracted from the image, and SIFT descriptors are computed for each case. The overlap is defined as half the length of the patch side. We use k-means clustering to build three visual vocabularies, one per patch size. The number of codewords in each vocabulary is set to 200.

[1]The number of histograms is $N_s \cdot \sum_{l=0}^{L-1} 2^{2l}$, where N_s is the number of patch scales and L is the number of levels in the pyramid.

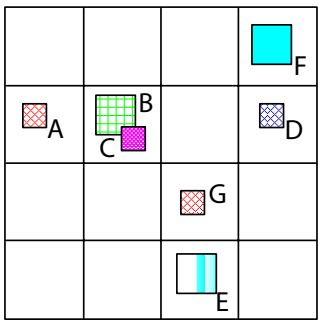


Figure 3. $\{A, B, C, D, E, F, G\}$ represent local patches and their layout on the 3rd level of the spatial pyramid (4×4 sub-regions). The texture filling each patch indicate different visual codewords. Our multi-scale local patches and spatial pyramid representation facilitates the extraction of relations among local patches. Relationships such as G is *above* E and F is *smoother* than D can be automatically mined. Note that one codeword may appear in a subregion for many times and one subregion usually contains patches of many different codewords.

3.2. Definition of Adjectives and Prepositions

Given our image representation, which captures patch appearances (based on a visual codebook) and spatial layout of patches (based on a spatial pyramid), we are now ready to define the concept of adjectives and prepositions. In our approach, the adjectives and prepositions can be constructed by simply considering particular spatial relationships among patches with specific appearances.

Figure 3 illustrates some adjectives and prepositions that can be represented with our method. The figure shows a spatial layout of the patches A, B, C, D, E, F, G in a 4×4 image partition. In the following, we describe some of the adjectives and prepositions that can be extracted from our image representation:

- `Above`(G, E): Patch G is above patch E.
- `Left`, `Near`(A, C): Patch A is on the left of C, and they are close to each other.
- `Left`, `Far`(A, D): Patch A is on the left of D, and they are far away from each other.
- `Larger`(B, C): Patch B is larger than C.
- `Brighter`(D, A): The intensity of patch D is brighter than the intensity of A.
- `Smoother`(F, E): The texture of patch F is smoother than the texture of patch E.

We can see that, based on the image representation of multi-scale local patches and the spatial pyramid, we can encode many different types of adjectives and prepositions, including relationships of texture, position, size, etc. Note

that the relationships that our method can represent are not limited to the ones listed above. In this work, we encode adjectives and prepositions in each spatial pyramid level independently, i.e. we do not consider the relations between patches from two different pyramid levels. In this method, the first level of the pyramid only encodes simple co-occurrence of image patches, with no information about their spatial layouts. As we move up in the pyramid level, more detailed spatial information is included in the adjectives and prepositions that are extracted.

Furthermore, instead of only considering relations among pairs of image patches, we consider the relation among multiple image patches simultaneously. This allows us to define more complex adjectives and prepositions, and even the relationships that combine adjectives and prepositions. For example, in Figure 3, we can encode the relations among $\{A, B, C\}$ simultaneously: (patches B and C are on the right of A) AND (B is larger than C).

4. Mining Discriminative Adjectives and Prepositions

4.1. Extraction of Adjectives and Prepositions As a Mining Problem

Having shown how the adjectives and prepositions are described, we are now ready to extract a set of discriminative adjectives and prepositions from the training images. Since we can represent not only adjectives and prepositions, but also the combination of them (Section 3.2), we denote the relationships that our approach can describe *Relationship Sets* (*RSets*). One RSet consists of several *Relationship Units* (*RUnits*); each unit indicates that a specific codeword appears in a specific image sub-region. A relationship that consists of m RUnits is called an m-RSet. Note that each RUnit is a 1-RSet.

An RSet $\mathcal{R}$ is discriminative for a class c if $\mathcal{R}$ has large occurrence scores on images belonging to class c, and has small occurrence scores on images of other classes. Considering an m-RSet $\mathcal{R} = \{R^1, \cdots, R^m\}$ and an image $\mathcal{I}$, the occurrence score s of $\mathcal{R}$ on $\mathcal{I}$ is computed by

$$s = \min\left\{\mathcal{I}(R^j),\ j = 1, \cdots, m\right\} \tag{1}$$

where $\mathcal{I}(R^j)$ is the number of occurrences of the RUnit R^j in image $\mathcal{I}$.

Given the occurrence score of each RSet, the discriminative ability of an RSet $\mathcal{R}$ for a class c is measured by two terms, the Support value $Supp(\mathcal{R}, c)$ and the Confidence value $Conf(\mathcal{R}, c)$. An RSet $\mathcal{R}$ is discriminative for a class c if both $Supp(\mathcal{R}, c)$ and $Conf(\mathcal{R}, c)$ are large. Let s_i denote the occurrence number of $\mathcal{R}$ in an image $\mathcal{I}_i$ with class label c_i. The support and confidence values of $\mathcal{R}$ for class c are computed by

$$Supp(\mathcal{R}, c) = \frac{\sum_{c_i=c} s_i}{\sum_{c_i=c} 1} \tag{2}$$

$$Conf(\mathcal{R}, c) = \frac{Supp(\mathcal{R}, c)}{\texttt{Avg}_{c' \neq c} Supp(\mathcal{R}, c')} \tag{3}$$

where $\texttt{Avg}_{c' \neq c} Supp(\mathcal{R}, c')$ indicates the average value of all support values $Supp(\mathcal{R}, c')$ where $c' \neq c$.

Intuitively, a large $Supp(\mathcal{R}, c)$ indicates that $\mathcal{R}$ generally has large occurrence numbers on images of class c, and a large $Conf(\mathcal{R}, c)$ implies small occurrence numbers of $\mathcal{R}$ on images of other classes. Therefore, in order to find discriminative RSets for a class, we want to find the RSets which have both a large support and a large confidence value on this class.

However, it is computationally expensive to evaluate support and confidence values for all RSets, because the number of relationship sets is extremely large. In the third pyramid level of our image representation, we have 16 image sub-regions and 600 codewords (200 for each local patch size). Therefore, only at this level, there are $600 \times 16 = 9600$ RUnits. Because we have to consider all possible combination of RUnits, the total number of potential RSets is 2^{9600}. In order to effectively explore this large space, we apply a data mining method [1], which enables us to obtain a set of discriminative RSets for each class efficiently.

Input: Support threshold T_{Supp} and confidence threshold T_{Conf}.
foreach *Class* **do**
  Scan all RUnits, select the RUnits with support values larger than T_{Supp} as 1-RSets;
  for p = *2 to* P **do**
    Generate candidate p-RSets based on the selected $(p-1)$-RSets;
    Scan all candidate p-RSets and remove the p RSets if the support values are smaller than T_{Supp};
    if *The number of p-RSets <2* **then**
      break;
    end
  end
  Scan all selected RSets, remove the RSets whose confidence values are smaller than T_{Conf}.
end

Algorithm 1: The Apriori mining algorithm. P is the total number of RUnits.

4.2. The Apriori Algorithm for Feature Mining

Apriori [1] is a data mining algorithm. The basic idea behind the Apriori algorithm is that, if an RSet has large support value, then any subset of that RSet must also have large support value. The Apriori [1] algorithm consists of two steps. The first step consists of several passes. The first pass finds all 1-RSets whose support values are larger than a threshold T_{Supp}. Then in a subsequent pass p, the algorithm first generates a set of candidate p-RSets based on the selected $(p-1)$-RSets, then computes the support value for each candidate p-RSet and remove the p-RSets whose support values are smaller than T_{Supp}. This step stops in the pass where no more RSets are generated. In the second step of the algorithm, confidence values are computed for all selected RSets, those with confidence values smaller than a threshold T_{Conf} will be removed. An overview of the Apriori algorithm is shown in Algorithm 1.

Implementation Details. In our implementation, we fix the value of T_{Conf} to 0.005, and use different values of T_{Supp} for different classes. The value of T_{Supp} for a class c is determined as follows. The support values and confidence values of all RUnits in the first pyramid level (without partitioning it into sub-regions) are computed. Then we select the RUnit with confidence values larger than $T_{Conf} = 0.005$, and rank their support values in a descending order. T_{Supp} for class c will be set as the 20-th largest value in this ranking. If the number of RUnits with larger confidence values is smaller than 20, then we set T_{Supp} to 0.005.

5. Scene Classification with RSets

In this section, we describe how to build classifiers for scene classification based on the mined RSets. For each image $\mathcal{I}$, we compute the occurrence number of all mined RSets within this image. Therefore each image can be represented as a D-dimensional vector, where D is the total number of mined RSets. Our classifier is an SVM with a histogram-intersection kernel that is trained on the D-dimensional vectors of all training images. Given two histograms $\mathcal{I}_1$ and $\mathcal{I}_2$, the histogram intersection kernel is computed by

$$\kappa = \sum_{d=1}^{D} \min\left(\mathcal{I}_1^d, \mathcal{I}_2^d\right) \tag{4}$$

where $\mathcal{I}_1^d$ is the value of the d-th bin in histogram $\mathcal{I}_1$.

6. Experiments

We carry out experiments on the fifteen scene categories data set from [8]. The dataset contains grayscale images of the following scene classes: highway, inside of cities, tall buildings, streets, suburb residence, forest, coast, mountain, open country, bedroom, kitchen, living room, office, industrial, and store. Each class has around $150 \sim 400$ images, and the average image size is approximately 300×250 pixels. This is one of the most complete natural scene category data set used in the literature so far.

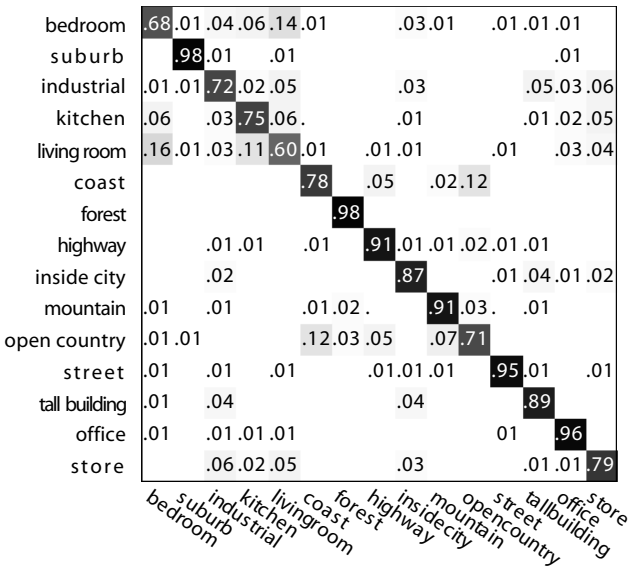


Figure 4. Confusion matrix of our approach on the scene category data set. Average classification results of 10 experiment runs are listed. The entry in the i-th row and j-th column is the percentage of images from class i that are classified as class j.

We use the same experiment setup as that in [8]. For each class, 100 images are randomly selected for training, and the remaining images are used for testing. We use the SVM^{light} package [7] to train our SVM classifier. We report performance accuracy as an average of 10 runs.

Table 1 shows the recognition accuracy of our approach. The 'Single-level' column summarizes the recognition accuracy when using patches and relations from a single pyramid level. The second column, labeled as 'Pyramid', summarizes the performance of our method when using patches and relationships extracted from all levels in the pyramid. We note that our method exhibits competitive performance to the state-of-the art results from [8]. We attribute such performance improvement to our stronger representation of spatial relationships among image patches.

Finally, recognition accuracy of each single level of our approach is also higher than the results in [8]. This is because the multi-scale local patch description can capture more information from the image.

The confusion matrix of our approach is shown in Figure 4. Our method tends to confuse visually similar scene classes, for example: bedroom vs. living room, kitchen vs. living room, and coast vs. open country.

Some mined relationships are shown in Figure 5. For each class, we only show the relationship of two codewords

Method	Accuracy	
Result in [2]	65.2	
Result in [8]:	Single-level	Pyramid
Level 1	74.8 ± 0.3	74.8 ± 0.3
Level 2	78.8 ± 0.4	80.1 ± 0.5
Level 3	79.7 ± 0.5	81.4 ± 0.5
Our Method:	Single-level	Pyramid
Level 1	75.3 ± 0.6	75.3 ± 0.6
Level 2	79.9 ± 0.4	82.6 ± 0.5
Level 3	80.4 ± 0.5	83.5 ± 0.6

Table 1. Recognition accuracy for different methods. The accuracy in [2] is measured in only 13 categories.

on the pyramid level that corresponds to 2×2 sub-regions. We observe that our approach can obtain some adjectives and prepositions with semantic meanings, e.g. sky above water in the 'coast' scene, sky above grass and trees in the 'suburb' scene.

7. Conclusion

This paper presents an approach to automatically mine adjectives and prepositions for natural scene recognition. Our method is based on an image representation that is built upon multi-scale image patches and a spatial pyramid. We apply a data mining approach to obtain a set of adjectives and prepositions, and combinations of adjectives and prepositions that are discriminative for the scene classification task. An interesting direction for future research would explore methods to automatically link the mined relationships to semantic descriptions in natural language.

8. Acknowledgements

The authors would like to thank Jia Li and Hao Su for the helpful discussions and comments. Li Fei-Fei is funded by a Microsoft Research fellowship and a Google award.

References

[1] R. Agrawal and R. Srikant. Fast algorithms for mining association rules. In *Proc. VLDB*, pages 487–499, 1994. 2, 4, 5

[2] L. Fei-Fei and P. Perona. A bayesian hierarchical model for learning natural scene categories. In *Proc. CVPR*, volume 2, pages 524–531, 2005. 1, 6

[3] P. F. Felzenszwalb and D. P. Huttenlocher. Efficient matching of pictorial structures. In *Proc. CVPR*, volume 2, pages 66–73, 2000. 2

[4] R. Fergus, P. Perona, and A. Zisserman. Object class recognition by unsupervised scale-invariant learning. In *Proc. CVPR*, volume 2, pages 264–271, 2003. 2

[5] K. Grauman and T. Darrell. The pyramid match kernel: Discriminative classification with sets of image features. In *Proc. ICCV*, volume 2, pages 1458–1465, 2005. 2

[6] A. Gupta and L. S. Davis. Beyond nouns: Exploiting prepositions and comparative adjectives for learning visual classifiers. In *Proc. ECCV*, volume 1, pages 16–29, 2008. 2

[7] T. Joachims. Making large-scale svm learning practical. Advances in kernel methods - support vector machines. B. Schölkopf, C. Burges, A. Smola (ed.), MIT press, 1999. 5

[8] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Proc. CVPR*, volume 2, pages 2169–2178, 2006. 2, 3, 5, 6

[9] D. G. Lowe. Object recognition from local scale-invariant features. In *Proc. ICCV*, volume 2, pages 1150–1157, 1999. 3

[10] T. Malisiewicz and A. A. Efros. Improving spatial support for objects via multiple segmentations. In *Proc. BMVC*, 2007. 2

[11] D. Parikh, C. L. Zitnick, and T. Chen. From appearance to context-based recognition: Dense labeling in small images. In *Proc. CVPR*, 2008. 2

[12] T. Quack, V. Ferrari, B. Leibe, and L. V. Gool. Efficient mining of frequent and distinctive feature configurations. In *Proc. ICCV*, 2007. 2

[13] A. Rabinovich, A. Vedaldi, C. Galleguillos, E. Wiewiora, and S. Belongie. Objects in context. In *Proc. ICCV*, 2007. 2

[14] S. Savarese, J. Winn, and A. Criminisi. Discriminative object class models of appearance and shape by correlatons. In *Proc. CVPR*, volume 2, pages 2033–2040, 2006. 2

[15] J. Sivic, B. C. Russell, A. A. Efros, A. Zisserman, and W. T. Freeman. Discovering objects and their location in images. In *Proc. ICCV*, volume 1, pages 370–377, 2005. 2

[16] J. Sivic, B. C. Russell, A. Zisserman, W. T. Freeman, and A. A. Efros. Unsupervised discovery of visual object class hierarchies. In *Proc. CVPR*, 2008. 2

[17] C. Wallraven, B. Caputo, and A. Graf. Recognition with local features: The kernel recipe. In *Proc. ICCV*, volume 1, pages 257–264, 2003. 1

[18] M. Weber, M. Welling, and P. Perona. Unsupervised learning of models for recognition. In *Proc. ECCV*, pages 18–32, 2000. 2

[19] J. Yuan, Y. Wu, and M. Yang. Discovery of collocation patterns: from visual words to visual phrases. In *Proc. CVPR*, 2007. 2

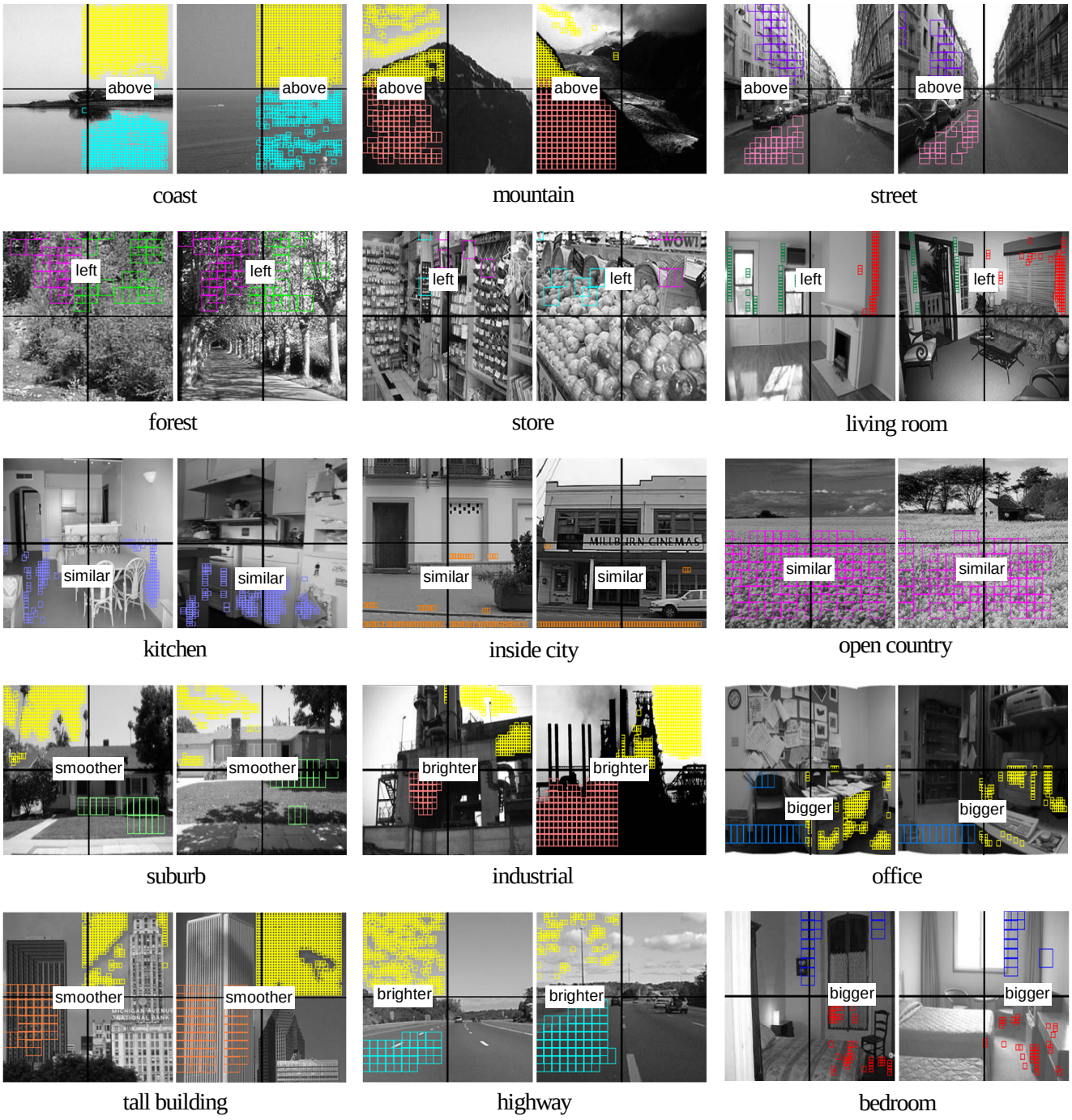


Figure 5. Examples of adjectives and prepositions that are inferred from the mined results. The image patches that are assigned to the same codeword are labeled with the same color. For example, in the "coast" images we show the relationship that yellow patches (sky) appear above cyan patches (water). In the "tall building" images we show the relationship that the texture of yellow patches (sky) is smoother than the orange patches (building).