Bag-of-visual-words and spatial extensions for land-use classification

Conference Paper · January 2010

DOI: 10.1145/1869790.1869829 · Source: DBLP

2 authors, including:

Shawn Newsam
University of California, Merced
74 PUBLICATIONS **1,928** CITATIONS

SEE PROFILE

Bag-Of-Visual-Words and Spatial Extensions for Land-Use Classification

Yi Yang and Shawn Newsam
Electrical Engineering & Computer Science
University of California at Merced
yyang6,snewsam@ucmerced.edu

ABSTRACT

We investigate bag-of-visual-words (BOVW) approaches to land-use classification in high-resolution overhead imagery. We consider a standard non-spatial representation in which the frequencies but not the locations of quantized image features are used to discriminate between classes analogous to how words are used for text document classification without regard to their order of occurrence. We also consider two spatial extensions, the established spatial pyramid match kernel which considers the absolute spatial arrangement of the image features, as well as a novel method which we term the spatial co-occurrence kernel that considers the relative arrangement. These extensions are motivated by the importance of spatial structure in geographic data.

The methods are evaluated using a large ground truth image dataset of 21 land-use classes. In addition to comparisons with standard approaches, we perform extensive evaluation of different configurations such as the size of the visual dictionaries used to derive the BOVW representations and the scale at which the spatial relationships are considered.

We show that even though BOVW approaches do not necessarily perform better than the best standard approaches overall, they represent a robust alternative that is more effective for certain land-use classes. We also show that extending the BOVW approach with our proposed spatial co-occurrence kernel consistently improves performance.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database Applications—*spatial databases and GIS*; I.5.4 [**Pattern Recognition**]: Applications; I.4.8 [**Image Processing and Computer Vision**]: Scene Analysis

Keywords

land-use classification, local invariant features, bag-of-visual-words

1. INTRODUCTION

ACM GIS '10 , November 2-5, 2010. San Jose, CA, USA

The paper investigates bag-of-visual-words (BOVW) approaches to land-use classification in high-resolution overhead imagery (we use the term overhead imagery to refer to both airborne and spaceborne imagery). BOVW approaches are motivated by document classification in text analysis and have been successfully applied to analyzing close-range imagery. We here present, however, what is, to the best of our knowledge, the first large scale application of BOVW approaches to land-use classification. We further investigate two spatial extensions, the established spatial pyramid match kernel, which considers the absolute spatial arrangement of an image, as well as a novel method which we term the spatial co-occurrence kernel that considers the relative arrangement. These extensions are motivated by the importance of spatial structure in geographic data.

We evaluate the methods using a large ground truth image dataset of 21 land-use classes. This manually labelled dataset is derived from images in the public domain and is made available for other researchers[1]. Besides comparing BOVW to standard approaches, namely classification based on color and texture features, we perform extensive evaluation of different configurations such as the size of the visual dictionaries used to derive the BOVW representations and the scale at which the spatial relationships are considered.

We conclude that even though BOVW approaches do not necessarily perform better than the best standard approaches overall, they represent a robust alternative that is more effective for certain land-use classes. And, since the current BOVW representation uses only the luminance information in an image, it can be combined with color information for further improvement. We also show that extending the BOVW approach with our proposed spatial co-occurrence kernel consistently improves performance.

2. BAG-OF-VISUAL-WORDS

This section describes the bag-of-visual-words (BOVW) approach to image representation. This method stems from text analysis wherein a document is represented by word frequencies without regard to their order. These frequencies are then used to perform document classification. Identifying the visual equivalent of a word is therefore necessary before the method can be applied to images. This is commonly done by extracting and quantizing local invariant features. We first discuss the motivation behind local invariant

[1]The labelled dataset can be downloaded from http://vision.ucmerced.edu/datasets.

features and then describe how they are transformed into visual words.

2.1 Local Invariant Features

Local invariant features have shown to be successful for a wide range of computer vision applications including wide-baseline stereo matching, object recognition, and category labelling. There are typically two steps in using local invariant features for image analysis. First, is a *detection* step which identifies interesting locations in the image usually according to some measure of saliency. These are termed interest points. Second, is to calculate a *descriptor* for each of the image patches centered at the detected locations. The following describes the desirable properties of the detection and descriptor components of local invariant features. These properties motivate their use for land-use classification.

Local The local property of the features makes their use robust to two common challenges in image analysis. First, they do not require the challenging preprocessing step of segmentation. The descriptors are not calculated for image regions corresponding to objects or parts of objects but instead for image patches at salient locations. Second, since objects are not considered as a whole, the features provide robustness against occlusion. They have been shown to reliably detect objects in cluttered scenes even when only portions of the objects are visible.

Invariance Local image analysis has a long history including corner and edge detection [9]. However, the success of the more recent approaches to local analysis is largely due to the invariance of the detection and descriptors to geometric and photometric image transformations. Note that it makes sense to discuss the invariance of both the detector and descriptor. An invariant detector will identify the same locations and image patches independent of a particular transformation. An invariant descriptor will remain the same. Often, the detection step estimates the transformation parameters necessary to normalize the image patch (to a canonical orientation and scale) so that the descriptor itself need not be completely invariant.

Geometric image transformations result from changes in viewing geometry and include translation, Euclidean (translation and rotation), similarity (translation, rotation, and uniform scaling), affine (translation, rotation, non-uniform scaling, and shear), and projective, the most general linear transformation in which parallel lines are not guaranteed to remain parallel. While affine invariant detectors have been developed [19], we choose a detector that is only invariant to similarity transformations for two reasons. First, remote sensed imagery is acquired at a relatively fixed viewpoint (overhead) which limits the amount of non-uniform scaling and shearing. Second, affine invariant detectors have been shown to perform worse than similarity invariant descriptors when the transformation is restricted to translation, rotation and uniform scaling [19]. Invariance to translation and scale is typically accomplished through scale-space analysis with automatic scale selection [14]. Invariance to rotation is typically accomplished by estimating the dominant orientation of the gradient of a scale-normalized image patch.

Photometric image transformations result from variations in illumination intensity and direction. Invariance is typically obtained in both the detector and descriptor by simply modelling the transformations as being linear and relying on changes in intensity rather absolute values. Utilizing intensity gradients accounts for the possible non-zero offset in the linear model and normalizing these gradients accounts for the possible non-unitary slope.

Robust yet distinctive The features should be robust to other transformations for which they are not designed to be invariant through explicit modelling. The detection and descriptor should not be greatly affected by modest image noise, image blur, discretization, compression artifacts, etc. Yet, for the features to be useful, the detection should be sufficiently sensitive to the underlying image signal and the descriptor sufficiently distinctive. Comprehensive evaluation has shown that local invariant features achieve this balance.

Density While detection is image dependent, it typically results in a large number of features. This density of features is important for robustness against occlusion as well as against missed and false detections. Of course, the large number of features that result from typical images present representation challenges. The histograms of quantized descriptors used in this work have shown to be an effective and efficient method for summarizing the features.

Efficient The extraction of local invariant features can be made computationally very efficient. This is important when processing large collections of images, such as is common in geographic image analysis, as well as for real-time applications. Real-time object detection has been demonstrated in prototype systems [22] as well as in commercial products [2].

2.2 SIFT Features

We choose David Lowe's Scale Invariant Feature Transform (SIFT) [15, 16] as the interest point detector and descriptor. While there are other detectors, such as the Harris-Laplace/Affine [19], Hessian-Laplace/Affine [19], Kadir and Brady's Saliency Detector [10]; other descriptors, such as shape context [4], steerable filters [5], PCA-SIFT [11], spin images [12], moment invariants [6], and cross-correlation; and other detector/descriptor combinations, such as Maximally Stable Extremal Regions (MSER) [18] and Speeded Up Robust Features (SURF) [3], we choose the SIFT detector and descriptor for the following reasons. First, the SIFT detector is translation, rotation, and scale invariant which is the level of invariance needed for our application. Second, an extensive comparison with other local descriptors found that the SIFT descriptor performed the best in an image matching task [20].

Interest point based image analysis, including SIFT, is a two-step process. First, a detection step locates points that are identifiable from different views. This process ideally locates the same regions in an object or scene regardless of viewpoint or illumination. Second, these locations are described by a descriptor that is distinctive yet invariant to viewpoint and illumination. SIFT-based analysis exploits image patches that can be found and matched under different imaging conditions.

2.2.1 SIFT Detector

The SIFT detection step is designed to find image regions that are salient not only spatially but also across different scales. Candidate locations are initially selected from local extrema in Difference of Gaussian (DoG) filtered images in scale space. The DoG images are derived by subtracting two Gaussian blurred images with different σ

$$D(x, y, \sigma) = L(x, y, k\sigma) - L(x, y, \sigma) \quad (1)$$

where $L(x, y, \sigma)$ is the image convolved with a Gaussian kernel with standard deviation σ, and k represents the different sampling intervals in scale space. Each point in the three dimensional DoG scale space is compared with its eight spatial neighbors at the same scale, and with its eighteen neighbors at adjacent higher and lower scales. The local maximum or minimum are further screened for minimum contrast and poor localization along elongated edges. The last step of the detection process uses a histogram of gradient directions sampled around the interest point to estimate its orientation. This orientation is used to align the descriptor to make it rotation invariant.

2.2.2 SIFT Descriptor

A SIFT descriptor is extracted from the image patch centered at each interest point. The size of this patch is determined by the scale of the corresponding extremum in the DoG scale space. This makes the descriptor scale invariant. The feature descriptor consists of histograms of gradient directions computed over a 4x4 spatial grid. The interest point orientation estimate described above is used to align the gradient directions to make the descriptor rotation invariant. The gradient directions are quantized into eight bins so the final feature vector has dimension 128 (4x4x8). This histogram-of-gradients descriptor can be roughly thought of a summary of the edge information in a scale and orientation normalized image patch centered at the interest point.

2.3 BOVW Representation

The SIFT detector, like most local feature detectors, results in a large number of interest points. This density is important for robustness but presents a representation challenge particularly since the SIFT descriptor features have 128 dimensions. We adopt a standard approach, termed bag-of-visual-words [23], to summarize the descriptors by quantizing and aggregating the features without regard to their location. The analogy to representing a text document by its word count frequencies is made possible by labelling each 128 dimension SIFT feature as a visual word. We apply standard k-means clustering to a large number of SIFT features to create a dictionary of visual words. This visual dictionary is then used to quantize the extracted features by simply assigning the label of the closest cluster centroid. The final representation for an image is the frequency counts or histogram of the labelled SIFT features

$$BOVW = [t_1, t_2, \ldots, t_M] \ , \tag{2}$$

where t_m is the number of occurrences of visual word m in the image and M is the dictionary size. To account for the difference in the number of interest points between images, the BOVW histogram is normalized to have unit L1 norm.

A BOVW representation can be used in kernel based learning algorithms, such as non-linear support vector machines, by computing the intersection between histograms. Given $BOVW^1$ and $BOVW^2$ corresponding to two images, the BOVW kernel is computed as:

$$K_{BOVW}(BOVW^1, BOVW^2) = \sum_{m=1}^{M} \min\left(BOVW^1(m), BOVW^2(m)\right). \tag{3}$$

The intersection kernel is a Mercer kernel which guarantees an optimal solution to kernel-based algorithms based on convex optimization such as nonlinear support vector machines.

3. SPATIAL EXTENSIONS TO BOVW

The BOVW approach does not consider the spatial locations of the visual words in an image (just as a BOW approach in text analysis does not consider where words appear in a document). One of the contributions of this paper is to explore spatial extensions to the BOVW approach for land-use classification. We are motivated by the obvious fact that spatial structure is important for geographic data and its analysis. As Walter Tobler stated in his so-called first law of geography in the early 1970's, all things are related, but nearby things are more related than distant things [24]. Since our objective is land-use classification in high-resolution overhead imagery of the earth's surface, this law motivates us to consider the spatial distribution of the visual words. In particular, we consider two extensions of the BOVW representation: the spatial pyramid match kernel and the spatial co-occurrence kernel. The spatial pyramid match kernel has shown to be successful for object and scene recognition in standard (non-overhead) imagery. It considers the *absolute* spatial arrangement of the visual words. By contrast, the spatial co-occurrence kernel, which we introduce here, considers the *relative* spatial arrangement.

3.1 Spatial Pyramid Match Kernel

The spatial pyramid match kernel (SPMK) was introduced by Lazebnik et al. in 2006 [13]. It is motivated by earlier work termed pyramid matching by Grauman and Darrell [7] on finding approximate correspondences between sets of points in high-dimensional feature spaces. The fundamental idea behind pyramid matching is to partition the feature space into a sequence of increasingly coarser grids and then compute a weighted sum over the number of matches that occur at each level of resolution. Two points are considered to match if they fall into the same grid cell and matched points at finer resolutions are given more weight than those at coarser resolutions. The SPMK applies this approach in the two-dimensional image space instead of a feature space; that is, it finds approximate spatial correspondence between sets of visual words in two images.

More specifically, suppose an image is partitioned into a sequence of spatial grids at resolutions $0, \ldots, L$ such that the grid at level l has 2^l cells along each dimension for a total of $D = 4^l$ cells. Let H_l^1 and H_l^2 be the histograms of visual words of two images at resolution l so that $H_l^1(i, m)$ and $H_l^2(i, m)$ are the counts of visual word m contained in grid cell i. Then, the number of matches at level l is computed as the histogram intersection:

$$I\left(H_l^1, H_l^2\right) = \sum_{i=1}^{D} \sum_{m=1}^{M} \min\left(H_l^1(i, m), H_l^2(i, m)\right). \tag{4}$$

Abbreviate $I\left(H_l^1, H_l^2\right)$ to I_l. Since the number of matches at level l includes all matches at the finer level $l + 1$, the number of new matches found at level l is $I_l - I_{l+1}$ for $l = 0, \ldots, L - 1$. Further, the weight associated with level l is set to $\frac{1}{2^{L-l}}$ which is inversely proportional to the cell size and thus penalizes matches found in larger cells. Finally,

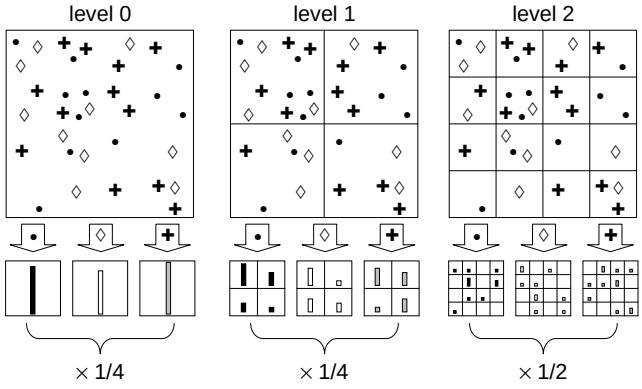


Figure 1: Toy example of a three-level spatial pyramid (adapted from [13]). The image has three visual words and is divided at three different levels of resolution. For each level, the number of words in each grid cell is counted. Finally, the spatial histogram is weighted according to equation 6.

the spatial pyramid match kernel for two images is given by:

$$K_L = I_L + \sum_{l=0}^{L-1} \frac{1}{2^{L-l}} (I_l - I_{l+1}) \quad (5)$$

$$= \frac{1}{2^L} I_0 + \sum_{l=1}^{L} \frac{1}{2^{L-l+1}} I_l. \quad (6)$$

The SPMK is a Mercer kernel. The SPMK is summarized in figure 1.

3.2 Spatial Co-occurrence Kernel

We take further motivation from early work on pixel-level characterization of land-use classes in overhead imagery from Haralick et al.'s seminal work [8] on gray level co-occurrence matrices (GLCM) and the set of 14 derived textural features which represents some of the earliest work on image texture. A GLCM provides a straightforward way to characterize the spatial dependence of pixel values in an image. We extend this to the spatial dependence of the visual words.

More formally, given an image I containing a set of n visual words $c_i \in C$ at pixel locations (x_i, y_i) and a binary spatial predicate ρ where $c_i \rho c_j \in \{0, 1\}$, we define the visual word co-occurrence matrix (VWCM) as

$$VWCM_\rho(u, v) = \|(c_i, c_j)| \, (c_i = u) \wedge (c_j = v) \wedge (c_i \rho c_j)\|. \quad (7)$$

That is, the VWCM is a count of the number of times two visual words satisfy the spatial predicate. The choice of the predicate ρ determines the nature of the spatial dependencies. While this framework provides the flexibility for variety of dependencies, we focus on proximity and, given a distance r, define ρ to be true if the two words appear within r pixels of each other:

$$c_i \rho_r c_j = \begin{cases} 1, & \text{if } \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2} \leq r; \\ 0, & \text{otherwise.} \end{cases} \quad (8)$$

The VWCMs computed in this paper thus represent the number of times pairs of words appear near to each other.

A fundamental challenge to using GLCMs or our proposed counterparts, VWCMs, is their size. For example, given a visual dictionary of size M, the VWCM has dimension $M \times M$. Even though a symmetric predicate such proximity results in a symmetric co-occurrence matrix, its size is still quadratic with respect to the dictionary, containing $M(M+1)/2$ entries. Haralick et al. therefore defined a set of 14 scalar quantities to summarize the GLCMs. We initially also summarized our VWCMs through six commonly used scalar quantities–entropy, maximum probability, correlation, contrast, energy, and homogeneity–but this did not prove to be effective for characterizing the spatial dependencies between visual words. We thus use the full co-occurrence matrix (up to its symmetry if applicable) and instead investigate smaller dictionaries.

Given two visual co-occurrence matrices $VWCM_\rho^1$ and $VWCM_\rho^2$ corresponding to images I^1 and I^2, we now compute the spatial co-occurrence kernel (SCK) as the intersection between the matrices

$$K_{SCK}(VWCM_\rho^1, VWCM_\rho^2) = \sum_{u,v \in C} \min(VWCM_\rho^1(u, v), VWCM_\rho^2(u, v)). \quad (9)$$

To account for differences between images in the number of pairs of codewords satisfying the spatial predicate, the matrices are normalized to have an L1 norm of one. Note that the SCK, as an intersection of two multidimensional counts, is also Mercer kernel and thus still guarantees an optimal solution in the learning stage of non-linear support vector machines.

3.3 SCK Combined With BOVW

While the proposed SCK can be used by itself, it can also serve as a spatial extension to the non-spatial BOVW representation. Specifically, given histograms $BOVW^1$ and $BOVW^2$ corresponding to two images, we compute the combined kernel as the sum of SCK and the intersection of the histograms

$$K_{SCK+BOVW}(\{VWCM_\rho^1, BOVW^1\}, \{VWCM_\rho^2, BOVW^2\}) = K_{SCK}(VWCM_\rho^1, VWCM_\rho^2) + K_{BOVW}(BOVW^1, BOVW^2). \quad (10)$$

Note that the visual dictionary used for the spatial co-occurrence matrices need not be the same as that used for the BOVW representation. We explore the effect of using different dictionaries in the experiments below. Again, since this combined kernel is a (positively weighted) sum of two Mercer kernels, it is itself a Mercer kernel. While it is possible to weight the spatial and non-spatial components of the combined kernel differently, we have so far not considered this and leave it for future work.

4. GLOBAL IMAGE DESCRIPTORS

We compare the BOVW representations with standard global image descriptors, namely color histograms and homogeneous texture.

4.1 Color Histograms

Color histogram descriptors are computed separately in three color spaces: RGB, hue lightness saturation (HLS), and CIE Lab. Each dimension is quantized into 8 bins for a total histogram feature length of 512. The histograms are normalized to have an L1 norm of one. This results

in three color histogram features: H_{RGB}, H_{HLS} and H_{Lab}. The intersection kernel is applied to the color histograms.

4.2 Homogeneous Texture

Homogeneous texture descriptors extracted using Gabor filters have proven effective for analyzing overhead imagery [21]. They were standardized in 2002 by the MPEG-7 Multimedia Content Description Interface [17] after they were shown to outperform other texture features in which one of the evaluation datasets consisted of high-resolution aerial imagery. We extract MPEG-7 compliant descriptors using a bank of Gabor filters tuned to five scales and six orientations. A 60 dimensional feature vector is then formed from the means and standard deviations of the 30 filters:

$$f_{MPEG7HT} = [\mu_{11}, \sigma_{11}, \mu_{12}, \sigma_{12}, \ldots, \mu_{1S}, \sigma_{1S}, \ldots, \mu_{RS}, \sigma_{RS}] \ , \quad (11)$$

where μ_{rs} and σ_{rs} are the mean and standard deviation of the output of the filter tuned to orientation r and scale s. A Gaussian radial basis function (RBF) kernel, again a Mercer kernel, is applied to the homogeneous texture descriptors.

5. DATASET

An extensive manually labelled ground truth dataset is used to perform quantitative evaluation. The dataset consists of images of 21 land-use classes selected from aerial orthoimagery with a pixel resolution of one foot. Large images were downloaded from the United States Geological Survey (USGS) National Map of the following US regions: Birmingham, Boston, Buffalo, Columbus, Dallas, Harrisburg, Houston, Jacksonville, Las Vegas, Los Angeles, Miami, Napa, New York, Reno, San Diego, Santa Barbara, Seattle, Tampa, Tucson, and Ventura. 100 images measuring 256×256 pixels were manually selected for each of the following 21 classes: agricultural, airplane, baseball diamond, beach, buildings, chaparral, dense residential, forest, freeway, golf course, harbor, intersection, medium density residential, mobile home park, overpass, parking lot, river, runway, sparse residential, storage tanks, and tennis courts. Note that we use the term land-use to refer to this set of classes even though they contain some land-cover and possibly object classes. These classes were selected because they contain a variety of spatial patterns, some homogeneous with respect to texture, some homogeneous with respect to color, others not homogeneous at all, and thus represent a rich dataset for our investigation.

Five samples of each class are shown in figure 2. The images downloaded from the National Map are in the RGB colorspace. The SIFT features underlying the BOVW representations and the homogeneous texture descriptors are extracted using the luminance channel.

6. EXPERIMENTS

The approaches are compared by performing multi-class classification. Classifiers are trained on a subset of the ground truth images and then applied to the remaining images. The classification rate is simply the percentage of the held-out images that are labelled correctly.

The premise behind our experiments is that once a classifier has been trained on a labelled dataset, it could be used to classify novel image regions. One of the benefits of a local feature based approach is that the regions need not be constrained to rectagonal or other regular shapes. Note that an image region need not be assigned one of the class labels and can instead be assigned a null-label if the classifier provides a confidence or similar score. Such is the case for support vector machines.

We use support vector machines (SVMs) to perform the classification. Multi-class classification is implemented using a set of binary classifiers and taking the majority vote. Non-linear SVMs incorporating the kernels described above are trained using grid-search for model selection. For the histogram intersection type kernels–BOVW, SPMK, SCK, BOVW+SCK, and color histogram–the only parameter is C, the penalty parameter of the error term. The RBF kernel used for homogeneous texture contains an addition width parameter γ. Five-fold cross-validation is performed in which the ground truth dataset is randomly split into five equal sets. The classifier is then trained on four of the sets and evaluated on the held-out set. The classification rate is the average over the five evaluations. Most results are presented as the average rate over all 21 classes. The SVMs were implemented using the LIBSVM package [1].

We compare a variety of difference configurations for the BOVW approaches:

- Different sized visual dictionaries for the BOVW and SPMK approaches: 10, 25, 50, 75, 100, 125, 150, 175, 200, 250, 300, 400, 500, 1000, and 5000.
- Different numbers of pyramid levels for the SPMK approach: 1 (essentially standard BOVW), 2, 3, and 4.
- Different sized visual dictionaries used to compute the co-occurrence matrices for the SCK approach: 10, 50, and 100.
- Different sized radii for the spatial predicate used to compute the co-occurrence matrices for the SCK approach: 20, 50, 100, and 150 pixels.

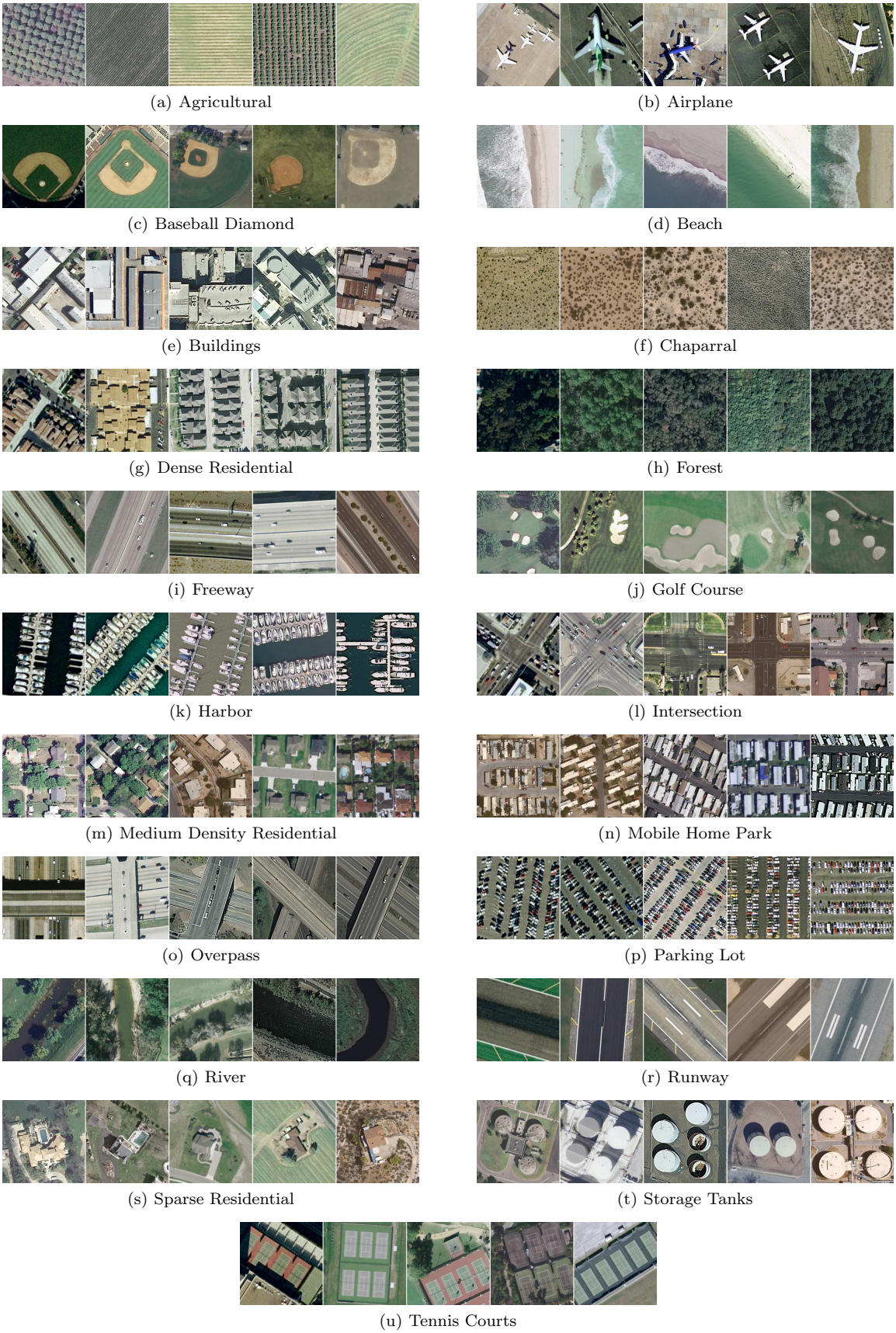
7. RESULTS

7.1 Overall

Table 1 shows the best average classification rate across all 21 classes for the different approaches. The best rates result from the following settings: for BOVW, a dictionary size of 1000; for SPMK, a dictionary size of 500 and a three level pyramid; for SCK, a co-occurrence dictionary size of 100 and a radius of 150; and for BOVW+SCK, a BOVW dictionary size of 1000, a co-occurrence dictionary of size 100, and a radius of 150. Overall, these results are impressive given that chance classification for a 21 class problem is only 4.76%. Interestingly, the average rate is similar for all the approaches with perhaps color histograms computed in the CIE Lab colorspace as the one outlier. Section 7.6 below compares the per-class rates which exhibit more variation between approaches.

Color histograms computed in the HLS colorspace perform the best overall, achieving a rate of 81.19%. This was somewhat unexpected because several of the classes exhibit significant inter-image color variation. But, the color and BOVW approaches are orthogonal in that the interest point descriptors are extracted using only the luminance channel so that a combined approach is likely to perform even better. This is possible future work.

The results from the BOVW approaches–BOVW, SPMK, and BOVW+SCK–for different sized dictionaries are compared visually in figure 3 and in tabular form in table 2.



(a) Agricultural

(b) Airplane

(c) Baseball Diamond

(d) Beach

(e) Buildings

(f) Chaparral

(g) Dense Residential

(h) Forest

(i) Freeway

(j) Golf Course

(k) Harbor

(l) Intersection

(m) Medium Density Residential

(n) Mobile Home Park

(o) Overpass

(p) Parking Lot

(q) River

(r) Runway

(s) Sparse Residential

(t) Storage Tanks

(u) Tennis Courts

Figure 2: The ground truth dataset contains 100 images from each of 21 land-use classes. Five samples from each class are shown above.

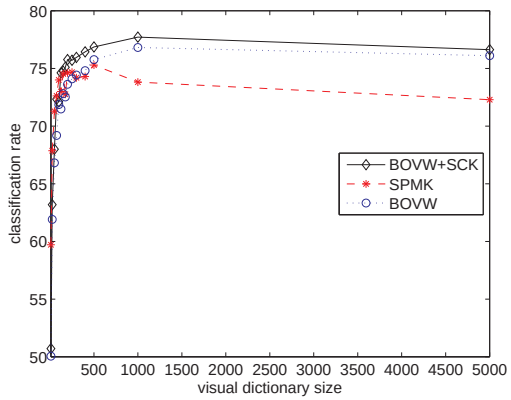


Figure 3: Comparison of BOVW, SPMK, and BOVW+SCK for different visual dictionary sizes. The size of visual dictionary used to derive the co-occurrence matrices for the SCK is as follows: 10 when the BOVW dictionary has size 10 or 25; 50 when the BOVW dictionary has size 50 or 75; and 100 otherwise. The radius used to derive the co-occurrences matrices is fixed at 150.

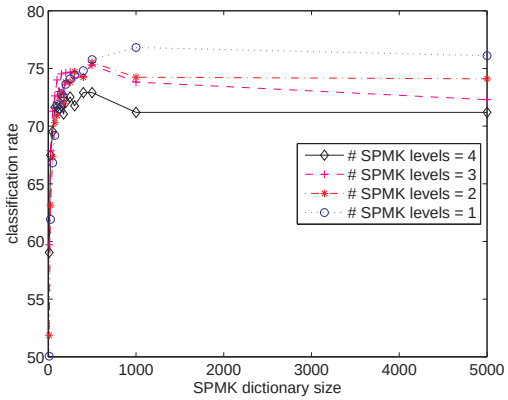


Figure 4: The effect of the number of levels used in the SPMK method.

The SPMK performs best for smaller dictionaries with 150 visual words or less. Smaller dictionaries correspond to a coarser quantization of the interest point feature space–that is, each visual word is less discriminative–so that the spatial arrangement of words is more important. But, as the dictionary size increases, the absolute spatial representation of SPMK actually leads to decreased performance over the non-spatial BOVW approach.

Significantly, BOVW+SCK outperforms BOVW for all dictionary sizes indicating that the relative spatial representation of SCK is complementary to the non-spatial information of BOVW. Therefore, a fundamental conclusion of this paper is that the SCK extension improves the BOVW approach for land-use classification. This is particulary true for smaller dictionary sizes which is significant from a computational viewpoint since the increased performance provided by the extension is almost equal to that which results from an order-of-magnitude increase in dictionary size for the non-spatial BOVW approach.

7.2 BOVW

As shown in table 2, a larger dictionary results in improved performance for the non-spatial BOVW up to around 1000 words. Very small dictionaries do not provide sufficient discrimination even though they might appeal from a computational and storage viewpoint. That performance decreases for very large dictionaries likely indicates that the visual words are too discriminative; that is, they are no longer robust to image perturbations caused by noise, blurring, discretization, etc. This decreases the likelihood that similar image patches are labelled as the same word.

7.3 SPMK

Figure 4 and table 3 provide further insight into the SPMK. Our results here confirm the findings of the originators of the method in that a spatial pyramid consisting of three levels tends to be optimum [13]. This remains true for dictionaries up to size 250 after which a single level pyramid which is the same as the non-spatial BOVW performs best. This indicates that the absolute spatial configuration of highly discriminative visual words is not effective for distinguishing the land-use classes.

7.4 SCK

Figure 5 shows the effect of the size of the dictionary and the radius of the spatial predicate used to compute the co-occurrence matrices for the SCK. The optimal configuration is a dictionary of size 100 and a radius of 150 pixels. We did not try dictionaries larger than 100 since the SCK representation grows quadratically with the number of visual words but the plots indicate that a slightly larger dictionary should increase performance. The results for the different radii indicate that longer range spatial interactions are significant for distinguishing the land-use classes. In particular, since our dataset has a ground resolution of one foot per pixel, the co-occurrence of visual words as far apart as 100 feet is discriminating. There seems to be little improvement past 100 feet though.

The SCK outperforms the BOVW for dictionaries of sizes 10, 50, and 100 for radii of 100 or 150 pixels (see table 3 for the BOVW values). It is unlikely, however, that this trend would continue for larger dictionaries (and it would be computationally and storage intensive) which motivates combining the BOVW and SCK approaches, possibly using different sized dictionaries for each.

7.5 BOVW+SCK

Section 7.1 above already showed that extending BOVW with SCK results in improved performance for all BOVW dictionary sizes. We now examine the effects of the SCK co-occurrence dictionary size and predicate radius on the combined method. Figure 6 and table 4 indicate that larger co-occurrence dictionary sizes result in improved performance although there is no clear winner between 50 and 100 visual words. Figure 7 and table 5 indicate that a larger spatial predicate radius results in improved performance again with little difference between radii of 100 and 150 pixels. All these observations are consistent with those of the SCK only method (see section 7.4 above) with the slight difference that a co-occurrence dictionary size of 100 does not consistently result in improved performance over a size of 50.

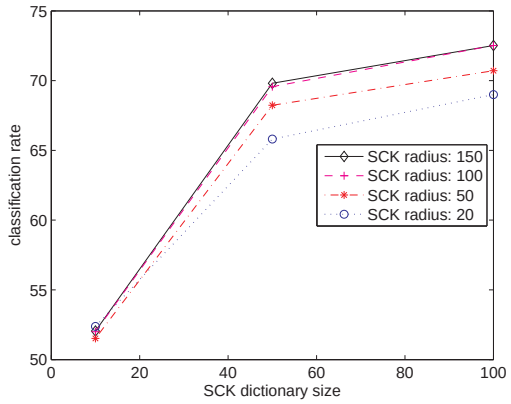


Figure 5: The effect of co-occurrence radius and dictionary size on the SCK method.

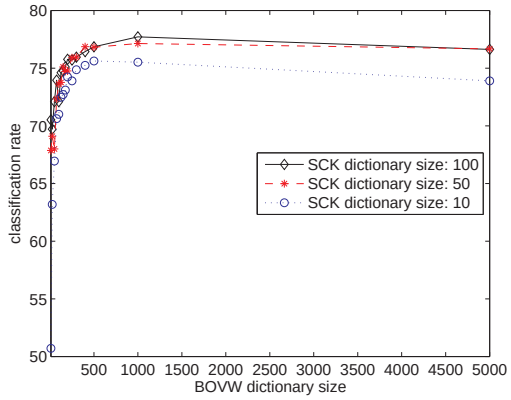


Figure 6: The effect of co-occurrence dictionary size on the BOVW+SCK method. The radius used to derive co-occurrence matrices is fixed at 150.

7.6 Per-Class Classification Rates

So far, we have only considered average classification rates over all classes. Figure 8 compares the per-class rates for the best configurations of the different methods. Not only is there significant variation between classes but there is also variation between methods within a class even though the methods do comparably when averaged over all classes. We summarize our observations as follows.

Easiest classes Chaparral, harbor, and parking lot, and to a certain extent forest, are the classes with the highest classification rates. These classes tend to be very homogeneous with respect to both color and texture. The variation between cars does result in color histogram features performing slightly worse than the other approaches on the parking lot class.

Most difficult classes Storage tanks and tennis courts, and to a degree the three residential classes, baseball diamond, and intersection, are the classes with the lowest classification rates. Indeed, these are the classes which have the most complex spatial arrangements as well as large inter-image variation. It is for many of these classes that color histograms outperform the other approaches since, as global features, they are invariant to spatial arrangement.

BOVW BOVW proves to be a robust "middle-of-the-pack" approach, never significantly outperforming nor underperforming the other techniques.

SPMK SPMK performs better than the other visual word approaches on the beach, building, runway, and tennis courts classes. This is due to the absolute spatial arrangement of these classes being important. Even though the coastline may be oriented differently in the beach images, for any particular orientation, the sand and the surf will be in the same image regions. The same is true for the runway images. SPMK performs the worst of all methods on the freeway class. While this class is similar to runway, this result is likely due to the different locations that the vehicles occur as well as the variation in the shoulders, medians, and road widths.

SCK SCK performs the best of all techniques on the forest and intersection classes. These are the classes for which relative spatial arrangement is important. It additionally performs better than SPMK on the agricultural, freeway and river classes, and when compared with SPMK also restricted to dictionary of size 100 (results not shown), this list also includes buildings, golf course, harbor, parking lot, and runway. These are the classes for which relative spatial arrangement is more important than absolute arrangement.

BOVW+SCK Extending the non-spatial BOVW approach using SCK maintains the robustness of BOVW while improving results for 12 of the 21 classes. If the common underlying BOVW component is restricted to a dictionary size of 100, BOVW+SCK outperforms BOVW for 16 of the 21 classes. This improvement is most significant for the beach and intersection classes. The SCK extension does result in a notable decrease in performance for the baseball diamond class (although this decrease is marginal for the smaller BOVW dictionary). This again supports one of the fundamental conclusions of this paper, that the SCK extension improves the BOVW approach for land-use classification.

Color As mentioned above, the performance of color histograms extracted in the HLS colorspace was somewhat unexpected. HLS histograms perform significantly better than the other methods on the baseball diamond, golf course, medium density residential, river, sparse residential, and storage tanks classes. This advantage over methods which only consider luminance is a result of the large intra-class homogeneity with respect to color. This advantage is compounded when HLS histograms are compared to the spatial methods since many of these classes have complex spatial arrangements often with large intra-class variation.

Texture Texture performs the best on the agricultural, airplane, freeway, and runway classes. These results are consistent with our previous use of homogeneous texture descriptors based on the outputs of Gabor filters for analyzing remote sensed imagery [21].

8. CONCLUSION

We described and evaluated BOVW and spatial extensions for land-use classification in high-resolution overhead imagery. While the BOVW-based approaches do not perform better overall than the best standard approach, they represent a robust alternative that is more effective for certain classes. We also proposed a novel spatial extension termed spatial co-occurrence kernel and showed that it con-

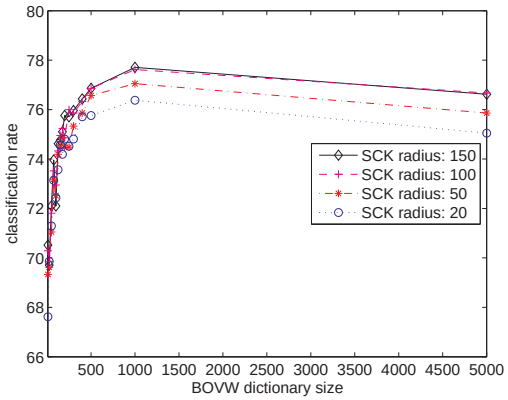


Figure 7: The effect of co-occurrence radius on the BOVW+SCK method. The size of the co-occurrence dictionary is fixed at 100.

sistently improves upon a BOVW baseline. Extensions of this work include further investigation into which classes interest point based approaches are the most appropriate for and integrating interest point and color analysis since they are complementary.

9. ACKNOWLEDGEMENTS

This work was funded in part by NSF grant IIS-0917069 and a Department of Energy Early Career Scientist and Engineer/PECASE award. Any opinions, findings, and conclusions or recommendations expressed in this work are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. The authors would like to thank the anonymous reviewers for their helpful comments.

10. REFERENCES

[1] LIBSVM–a library for support vector machines. http://www.csie.ntu.edu.tw/ cjlin/libsvm/.

[2] Snaptell - visual product search. http://snaptell.com/.

[3] H. Bay, T. Tuytelaars, and L. V. Gool. SURF: Speeded up robust features. In *European Conference on Computer Vision*, 2006.

[4] S. Belongie, J. Malik, and J. Puzicha. Shape matching and object recognition using shape contexts. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 24(4):509–522, 2002.

[5] W. T. Freeman and E. H. Adelson. The design and use of steerable filters. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 13(9):891–906, 1991.

[6] L. J. V. Gool, T. Moons, and D. Ungureanu. Affine/photometric invariants for planar intensity patterns. In *European Conference on Computer Vision*, 1996.

[7] K. Grauman and T. Darrell. The pyramid match kernel: Discriminative classification with sets of image features. In *IEEE International Conference on Computer Vision*, 2005.

[8] R. M. Haralick, K. Shanmugam, and I. Dinstein. Texture features for image classification. *IEEE Transactions on Systems, Man, and Cybernetics*, 3:610–621, 1973.

[9] C. Harris and M. Stephens. A combined corner and edge detector. In *Proceedings of The Fourth Alvey Vision Conference*, 1988.

[10] T. Kadir, A. Zisserman, and M. Brady. An affine invariant salient region detector. In *European Conference on Computer Vision*, 2004.

[11] Y. Ke and R. Sukthankar. PCA-SIFT: a more distinctive representation for local image descriptors. In *IEEE International Conference on Computer Vision and Pattern Recognition*, 2004.

[12] S. Lazebnik, C. Schmid, and J. Ponce. Sparse texture representation using affine-invariant neighborhoods. In *IEEE International Conference on Computer Vision and Pattern Recognition*, 2003.

[13] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *IEEE International Conference on Computer Vision and Pattern Recognition*, 2006.

[14] T. Lindeberg. Feature detection with automatic scale selection. *International Journal of Computer Vision*, 30(2):79–116, 1998.

[15] D. G. Lowe. Object recognition from local scale-invariant features. In *IEEE International Conference on Computer Vision*, 1999.

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

[17] B. S. Manjunath, P. Salembier, and T. Sikora, editors. *Introduction to MPEG7: Multimedia Content Description Interface*. John Wiley & Sons, 2002.

[18] J. Matas, O. Chum, M. Urban, and T. Pajdla. Robust wide baseline stereo from maximally stable extremal regions. In *British Machine Vision Conference*, 2002.

[19] K. Mikolajczyk and C. Schmid. Scale & affine invariant interest point detectors. *International Journal of Computer Vision*, 60(1):63–86, 2004.

[20] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 27(10):1615–1630, 2005.

[21] S. Newsam, L. Wang, S. Bhagavathy, and B. S. Manjunath. Using texture to analyze and manage large collections of remote sensed image and video data. *Journal of Applied Optics: Information Processing*, 43(2):210–217, 2004.

[22] D. Nister and H. Stewenius. Scalable recognition with a vocabulary tree. In *IEEE International Conference on Computer Vision and Pattern Recognition*, 2006.

[23] J. Sivic and A. Zisserman. Video Google: A text retrieval approach to object matching in videos. In *IEEE International Conference on Computer Vision*, 2003.

[24] W. Tobler. A computer movie simulating urban growth in the Detroit region. *Economic Geography*, 46(2):234–240, 1970.

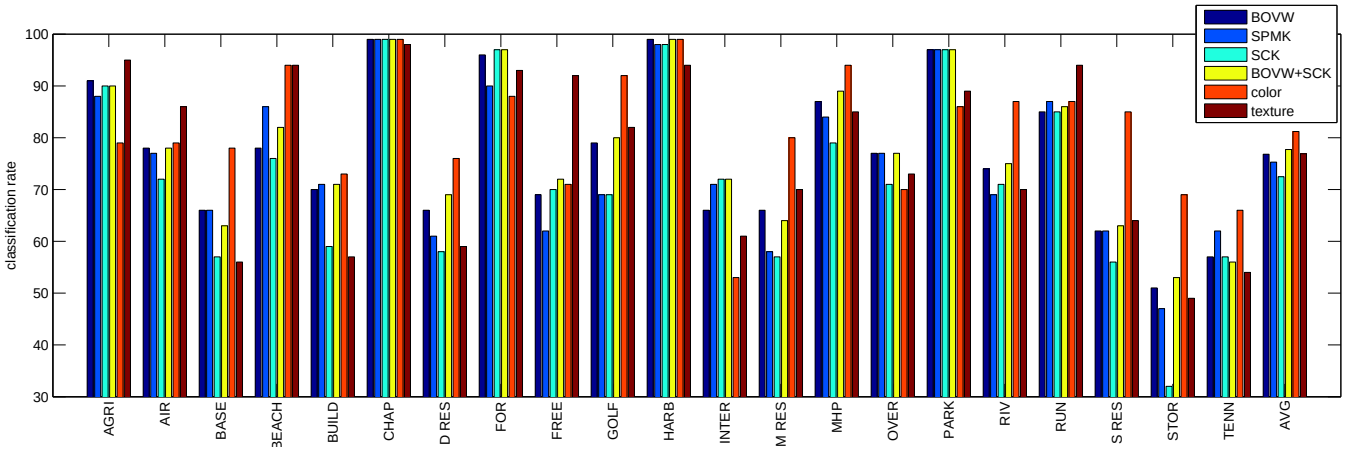


Figure 8: Per-class classification rates corresponding to the optimal configuration for each method. BOVW uses a dictionary size of 1000. SPMK uses a dictionary size of 500 with three levels. SCK uses a co-occurrence dictionary size of 100 and radius of 150. BOVW+SCK uses a BOVW dictionary size of 1000 and a co-occurrence dictionary size of 100 and radius of 150. The color histograms are computed in the HLS colorspace.

Table 1: Best classification rates for different approaches. See text for details.

	BOVW	SPMK	SCK	BOVW+SCK	Color–RGB	Color–HLS	Color–Lab	Texture
Rate	76.81	75.29	72.52	77.71	76.71	81.19	66.43	76.91

Table 2: Classification rates for BOVW, SPMK, and BOVW+SCK for different visual dictionary sizes. The size of visual dictionary used to derive the co-occurrence matrices for the SCK is as follows: 10 when the BOVW dictionary has size 10 or 25; 50 when the BOVW dictionary has size 50 or 75; and 100 otherwise. The radius used to derive the co-occurrences matrices is fixed at 150.

	10	25	50	75	100	125	150	175	200	250	300	400	500	1000	5000
BOVW	50.05	61.91	66.81	69.19	71.86	71.48	72.81	72.52	73.62	74.10	74.43	74.81	75.76	76.81	76.10
SPMK	59.71	67.86	71.29	72.62	74.00	73.00	74.52	72.76	74.62	74.67	74.19	74.29	75.29	73.81	72.29
BOVW+SCK	50.71	63.19	68.00	72.33	72.10	74.62	74.76	75.10	75.76	75.71	75.95	76.43	76.86	77.71	76.62

Table 3: The effect of the number of levels used in the SPMK method.

	10	25	50	75	100	125	150	175	200	250	300	400	500	1000	5000
1	50.05	61.91	66.81	69.19	71.86	71.48	72.81	72.52	73.62	74.10	74.43	74.81	75.76	76.81	76.10
2	51.86	63.14	67.33	70.29	70.91	72.52	72.81	71.95	73.81	73.81	74.71	74.24	75.52	74.24	74.10
3	59.71	67.86	71.29	72.62	74.00	73.00	74.52	72.76	74.62	74.67	74.19	74.29	75.29	73.81	72.29
4	59.05	67.48	69.52	71.62	71.62	71.52	71.76	71.05	72.05	72.52	71.76	72.90	72.90	71.19	71.19

Table 4: The effect of co-occurrence dictionary size on the BOVW+SCK method. The rows correspond to the size of the dictionary used to derive the co-occurrence matrices. The columns correspond to the size of the dictionary for the BOVW component. The radius used to derive the co-occurrence matrices is fixed at 150.

	10	25	50	75	100	125	150	175	200	250	300	400	500	1000	5000
10	50.71	63.19	66.95	70.62	71.00	72.48	72.71	73.10	74.24	73.90	74.86	75.24	75.62	75.52	73.90
50	67.86	69.10	68.00	72.33	73.62	73.76	75.10	74.71	74.76	75.90	75.95	76.86	76.81	77.14	76.67
100	70.52	69.71	72.10	73.95	72.10	74.61	74.76	75.10	75.76	75.71	75.95	76.43	76.85	77.71	76.62

Table 5: The effect of co-occurrence radius on the BOVW+SCK method. The rows correspond to the radius used to derive the co-occurrence matrices. The columns correspond to the size of the dictionary for the BOVW component. The size of the co-occurrence dictionary is fixed at 100.

	10	25	50	75	100	125	150	175	200	250	300	400	500	1000	5000
20	67.62	69.86	71.29	73.14	72.43	73.57	74.57	74.19	74.81	74.52	74.81	75.71	75.76	76.38	75.05
50	69.33	69.62	71.05	73.19	72.52	74.19	74.48	74.86	74.48	74.48	75.33	75.86	76.57	77.05	75.86
100	70.29	69.86	71.81	73.52	72.95	74.33	74.95	75.24	74.86	76.00	75.86	76.33	76.86	77.62	76.67
150	70.52	69.71	72.10	73.96	72.10	74.62	74.76	75.10	75.76	75.71	75.96	76.43	76.86	77.71	76.62