Spatial Pyramid Co-occurrence for Image Classification

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

Abstract

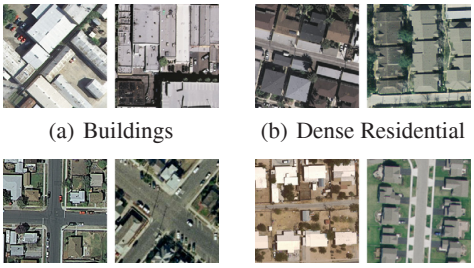
We describe a novel image representation termed spatial pyramid co-occurrence *which characterizes both the photometric and geometric aspects of an image. Specifically, the co-occurrences of visual words are computed with respect to spatial predicates over a hierarchical spatial partitioning of an image. The representation captures both the absolute and relative spatial arrangement of the words and, through the choice and combination of the predicates, can characterize a variety of spatial relationships.*

Our representation is motivated by the analysis of overhead imagery such as from satellites or aircraft. This imagery generally does not have an absolute reference frame and thus the relative spatial arrangement of the image elements often becomes the key discriminating feature. We validate this hypothesis using a challenging ground truth image dataset of 21 land-use classes manually extracted from high-resolution aerial imagery. Our approach is shown to result in higher classification rates than a non-spatial bag-of-visual-words approach as well as a popular approach for characterizing the absolute spatial arrangement of visual words, the spatial pyramid representation of Lazebnik et al. [7]. While our primary objective is analyzing overhead imagery, we demonstrate that our approach achieves state-of-the-art performance on the Graz-01 object class dataset and performs competitively on the 15 Scene dataset.

1. Introduction

Local invariant features have proven effective for a range of computer vision problems over the last decade. These features characterize the photometric aspects of an image while allowing for robustness against variations in illumination and noise. The geometric aspects of an image can further be characterized by considering the spatial arrangement of the local features.

This paper proposes a novel image representation termed *spatial pyramid co-occurrence* which characterizes both the photometric and geometric aspects of an image. Specifically, the co-occurrences of visual words–quantized local invariant features–are computed with respect to spatial predicates over a hierarchical spatial partitioning of an image. The local co-occurrences combined with the global partitioning allows the proposed approach to capture both the relative and absolute layout of an image. This is one of the salient aspects of spatial pyramid co-occurrence.



(a) Buildings (b) Dense Residential

(c) Intersection (d) Med. Density Res.

Figure 1. Our primary focus is on analyzing overhead imagery which generally does not have an absolute reference frame. The relative spatial arrangement of the image elements often becomes the key discriminating feature as demonstrated in the four land-use classes above.

Another salient aspect of the proposed approach is that it is general enough to characterize a variety of spatial arrangements. We give examples of spatial predicates which constrain the distances between pairs of visual words, the relative orientation between pairs of words, or both.

We are motivated by the problem of analyzing overhead imagery such as from satellites or aircraft. This imagery generally does not have an absolute reference frame and thus the relative spatial arrangement of the image elements often becomes the key discriminating feature. See, for example, the images of different land-use classes in figure 1.

We evaluate our approach using a novel ground truth image dataset of 21 land-use classes manually extracted from publicly available high-resolution overhead imagery. This dataset is one of the first of its kind and will be made available for other researchers[1]. Our approach is shown to result in higher classification rates on the land-use dataset than a non-spatial bag-of-visual-words approach as well as a popu-

[1]The dataset is available at http://vision.ucmerced.edu/datasets.

2011 IEEE International Conference on Computer Vision
978-1-4577-1102-2/11/$26.00

lar approach for characterizing the absolute spatial arrangement of visual words, the spatial pyramid representation of Lazebnik et al. [7].

We perform a thorough evaluation of the effects of different configurations of our approach such as the size of the visual dictionary and the specificity of the spatial predicate. We report interesting findings like the fact that smaller visual dictionaries become preferable for the co-occurrence component of our representation as the spatial predicate becomes more specialized. This somewhat counter-intuitive result has important implications for the computational complexity of our representation.

Finally, even though our primary objective is analyzing overhead imagery, we demonstrate that our approach achieves state-of-the-art performance on the Graz object class evaluation dataset and performs competitively on the 15 Scene evaluation dataset.

2. Related Work

The broader context of our work is bag-of-visual-words (BOVW) [3, 13] approaches to image classification. These approaches quantize local invariant image descriptors using a visual dictionary typically constructed through k-means clustering. The set of visual words is then used to represent an image regardless of their spatial arrangement similar to how documents can be represented as an unordered set of words in text analysis. The quantization of the often high-dimensional local descriptors provides two important benefits: it provides further invariance to photometric image transformations, and it allows compact representation of the image such as through a histogram of visual word counts and/or efficient indexing through inverted files. The size of the visual dictionary used to quantize the descriptors controls the tradeoff between invariance/efficiency and discriminability.

Lazebnik et al. [7] was one of the first works to address the lack of spatial information in the BOVW representation. Their spatial pyramid representation was motivated by earlier work termed pyramid matching by Grauman and Darrell [4] 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 spatial pyramid representation of Lazebnik et al. applies this approach in the two-dimensional image space instead of the feature space; that is, it finds approximate spatial correspondences between sets of visual words in two images.

The spatial pyramid representation characterizes the absolute location of the visual words in an image. Saverese et al. [12] propose a model which instead characterizes the relative locations. Motivated by earlier work on using correlograms of quantized colors for indexing and classifying images [6], they use correlograms of visual words to model the spatial correlations between quantized local descriptors. The correlograms are three dimensional structures which in essence record the number of times two visual words appear at a particular distance from each other. Correlogram elements corresponding to a particular pair of words are quantized to form correlations. Finally, images are represented as histograms of correlations and classified using nearest neighbor search against exemplar images. One challenge of this approach is that the quantization of correlograms to correlations can discard the identities of associated visual word pairs and thus may diminish the discriminability of the local image features.

Ling and Soatto [8] also characterize the relative locations of visual words. Their proximity distribution representation is a three dimensional structure which records the number of times a visual word appears within a particular number of nearest neighbors of another word. It thus captures the distances between words based on ranking and not absolute units. A corresponding proximity distribution kernel is used for classification in a support vector machine (SVM) framework. However, since proximity kernels are applied to the whole image, distinctive local spatial distributions of visual words may be overshadowed by global distributions.

Liu et al. [9] extend the BOVW framework by calculating spatial histograms where the co-occurrences of local features are calculated in circular regions of varying distances. However, the spatial histograms are only extracted for select visual words determined through an additional feature selection algorithm. Also, the spatial histograms are generated by averaging the counts of co-occurrences throughout the entire image and thus may also fail to capture distinctive local spatial arrangements.

Our proposed spatial pyramid co-occurrence differs from the above approaches in the following ways:

- It characterizes both the absolute and relative spatial layout of an image.
- It can characterize a greater variety of local spatial arrangements through the underlying spatial predicate. For example, combined proximity and orientation predicates can capture the general spatial distribution of visual words as well as the shape of local regions.
- The approach is simple in that it does not require learning a generative or other form of model.
- The representation can be easily combined with other representations such as a non-spatial bag-of-visual-words. And, since the representations are fused late,

visual dictionaries of different sizes can be used for the spatial (co-occurrence) and non-spatial components of the combined representation. This allows the non-spatial component to leverage the increased discriminability of the larger dictionary while limiting the computational costs associated with storing and comparing the co-occurrence structures.

3. Methods

Spatial pyramid co-occurrence characterizes both the absolute and relative spatial arrangement of visual words in an image. After stating our assumptions and describing the non-spatial BOVW representation, we review the spatial pyramid representation of Lazebnik et al. [7] since the proposed method uses the same hierarchical decomposition of an image. We then describe the proposed approach.

3.1. Assumptions

We assume each image I contains a set of N visual words c_i at pixel locations (x_i, y_i) where each word has been assigned a discrete label $c_i \in [1...M]$ from a visual dictionary containing M entries. The locations of the visual words could either be determined using a (dense) grid or an interest-point/saliency detector. We use Lowe's scale invariant feature transform (SIFT) detector [10] in the experiments below. Local invariant features are extracted at these locations and quantized into a discrete set of labels using a codebook typically generated by applying k-means clustering to a large, random set of features. We also use Lowe's SIFT descriptor [10] in the experiments below.

3.2. BOVW Representation

The non-spatial BOVW representation simply records the visual word occurrences in an image. It is typically represented as a histogram

$$BOVW = [t_1, t_2, \ldots, t_M] \ ,$$

where t_m is the number of occurrences of visual word m. To account for the difference in the number of visual words between images, the BOVW histogram is typically 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 $BOVW1$ and $BOVW2$ corresponding to two images, the BOVW kernel is computed as:

$$K_{BOVW}(BOVW1, BOVW2) = \sum_{m=1}^{M} \min\left(BOVW1(m), BOVW2(m)\right).$$

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

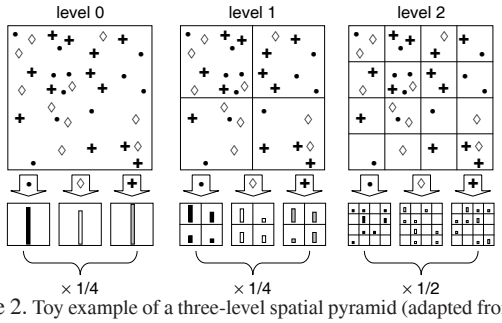


Figure 2. Toy example of a three-level spatial pyramid (adapted from [7]). The image has three visual words and is divided at three different levels of resolution. For each level, the number of words in each grid cell is counted. Finally, the spatial histogram is weighted according to equation 1.

3.3. Spatial Pyramid

The spatial pyramid representation of Lazebnik et al. [7] partitions an image 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. A BOVW histogram is then computed separately for each cell in the multiresolution grid. Specifically, $H_l(k, m)$ is the count of visual word m contained in grid cell k at level l. This representation is summarized in figure 2.

A spatial pyramid match kernel (SPMK) is derived as follows. Let $H1_l$ and $H2_l$ be the histograms of two images at resolution l. Then, the number of matches at level l is computed as the histogram intersection:

$$I\left(H1_l, H2_l\right) = \sum_{k=1}^{D} \sum_{m=1}^{M} \min\left(H1_l\left(k, m\right), H2_l\left(k, m\right)\right).$$

Abbreviate $I\left(H1_l, H2_l\right)$ 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, the SPMK for two images is given by:

$$K_{SPMK} = I_L + \sum_{l=0}^{L-1} \frac{1}{2^{L-l}} \left(I_l - I_{l+1}\right). \qquad (1)$$

The SPMK is also a Mercer kernel [7].

3.4. Spatial Co-occurrence

Spatial co-occurrence of visual words is motivated by Haralick et al.'s seminal work [5] on gray level co-occurrence matrices (GLCM) which is 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 visual words.

Formally, given an image I containing a set of N visual words c_i at pixel locations (x_i, y_i) and a binary spatial pred-

icate ρ where $c_i \rho c_j \in \{T, F\}$, we define the visual word co-occurrence matrix (VWCM) as

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

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. This framework can support a variety of dependencies such as the two visual words needing to be within a certain distance of each other, to have the same orientation, etc. We describe a number of predicates in the experiments section.

We derive a spatial co-occurrence kernel (SCK) as follows. Given two visual co-occurrence matrices $VWCM1_\rho$ and $VWCM2_\rho$ corresponding to two images, the SCK is computed as the intersection between the matrices

$$K_{SCK_\rho}(VWCM1_\rho, VWCM2_\rho) = \sum_{u,v \in M} \min(VWCM1_\rho(u,v), VWCM2_\rho(u,v)).$$

To account for differences in the number of pairs of codewords satisfying the spatial predicate between images, the matrices are normalized to have an L1 norm of one. The SCK, as an intersection of two multidimensional counts, is also Mercer kernel.

3.5. Combining Multiple Spatial Predicates

Multiple binary spatial predicates can be combined as follows. Given co-occurrence matrices $VWCM1_{\rho_A}(u,v)$ and $VWCM2_{\rho_A}(u,v)$ corresponding to predicate ρ_A for two images, and co-occurrence matrices $VWCM1_{\rho_B}(u,v)$ and $VWCM2_{\rho_B}(u,v)$ corresponding to predicate ρ_B for the same two images, a single SCK is computed as the sum of the individual SCKs

$$K_{SCK_{\rho_A+\rho_B}} = K_{SCK_{\rho_A}}(VWCM1_{\rho_A}, VWCM2_{\rho_A}) + K_{SCK_{\rho_B}}(VWCM1_{\rho_B}, VWCM2_{\rho_B}).$$

This too is a Mercer kernel. While it is possible to weight the components corresponding to the two predicates differently, we have so far not considered this and leave it for future work.

3.6. Spatial Pyramid Co-occurrence

We now describe the main contribution of this paper, *spatial pyramid co-occurrence*. Again, 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. The spatial co-occurrence of visual words is then computed separately for each cell in the multiresolution grid. Specifically, given a binary spatial predicate ρ, compute

$$VWCM^l_\rho(k,u,v) = \|(c_i, c_j)|\ (c_i = u) \wedge (c_j = v) \wedge (c_i \rho c_j)\|$$

where the visual words c_i are restricted to those in grid cell k at pyramid level l.

A spatial pyramid co-occurrence kernel (SPCK) corresponding to the spatial pyramid co-occurrences for two images $VWCM1_\rho$ and $VWCM2_\rho$ is then computed as

$$K_{SPCK}(VWCM1_\rho, VWCM2_\rho) = \sum_{l=0}^{L} w_l \sum_{k=1}^{D} \sum_{u,v \in M} \min(VWCM1^l_\rho(k,u,v), VWCM2^l_\rho(k,u,v))$$

where the weights w_l are chosen so that the sum of intersections has the same maximum achievable value for each level; e.g., $w_l = 1/4^l$. As a sum of intersections, the SPCK is a Mercer kernel.

Note that the spatial pyramid co-occurrence representation captures both the absolute and relative spatial arrangements of the visual words. The pyramid decomposition characterizes the absolute locations through the hierarchical gridding of the image and the VLCMs characterize the relative arrangements within the individual grid cells.

Multiple binary spatial predicates can again be combined by summing the SPCKs corresponding to the individual predicates.

3.7. Extended SPCK

The SPCK and the non-spatial BOVW representations are complementary and so it is natural to consider combining them. We thus form an extended SPCK representation, termed SPCK+, as the sum of the individual kernels:

$$K_{SPCK+}(\{VWCM1_\rho, BOVW1\}, \{VWCM2_\rho, BOVW2\}) = K_{SPCK}(VWCM1_\rho, VWCM2_\rho) + K_{BOVW}(BOVW1, BOVW2).$$

This sum is a Mercer kernel. We have not considered different weights for the the spatial and non-spatial components of the combined kernel and leave this too for future work.

Note that since the spatial and non-spatial components of our representation are fused late, *the visual dictionary used to derive the spatial co-occurrence matrices need not be the same as that used to derive the BOVW histograms*. Indeed, the experiments below show that smaller co-occurrence dictionaries are preferable for SPCK+ as the spatial predicates become more specialized. This helps reduce the computational complexity of the proposed approach.

Since SPCK and SPMK are also complementary in how they characterize spatial dependencies, we also consider a second extended SPCK representation, termed SPCK++, as the sum of the SPCK and SPMK kernels:

$$K_{SPCK++} = K_{SPCK} + K_{SPMK}. \quad (2)$$

3.8. Computational Complexity

We compare the computational costs of BOVW, SMPK, and SPCK in terms of the sizes of the different representations and the operations required to evaluate the kernels.

For a dictionary of size M, the BOVW representation has size M and evaluating the BOVW kernel requires M min computations (plus $M-1$ additions). For the same sized dictionary, an SPMK representation with levels $0,\ldots,L$ has size

$$S_{SPMK} = \sum_{l=0}^{L} \sum_{k=1}^{l^4} M$$

and evaluating the SPMK kernel requires the same number of min computations. For $L=2$, $S_{SPMK} = 21M$.

A VLCM corresponding to a co-occurrence dictionary of size N has N^2 entries (this reduces to $N(N+1)/2$ unique entries for symmetric spatial predicates such as those used in the experiments below). So, a SPCK representation with levels $0,\ldots,L$ has size

$$S_{SPCK} = \sum_{l=0}^{L} \sum_{k=1}^{l^4} N^2$$

and evaluating the SPCK kernel requires the same number of min computations. For $L=2$, $S_{SPCK} = 21N^2$. In the case where $N \leq \sqrt{M}$, the computational complexity of SPCK in terms of storage and kernel-evaluation is $O(M)$, the same as for BOVW and SPMK. This remains true when combining multiple spatial predicates. This is significant with respect to the finding in the experiments below that greatly reduced co-occurrence dictionaries are sufficient or even optimal for the extended SPCK representations.

4. Experiments and Results

We evaluate our proposed spatial pyramid co-occurrence representation on three datasets: 1) a novel dataset of land-use classes in high-resolution overhead imagery, 2) the publicly available Graz-01 object class evaluation dataset, and 3) the publicly available 15 Scene evaluation dataset.

4.1. Spatial Predicates

We consider two types of spatial predicates: *proximity* predicates which characterize the distance between pairs of visual words, and *orientation* predicates which characterize the relative orientations of pairs of visual words.

Proximity Since our primary goal is to analyze overhead imagery, and, according to Tobler's first law of geography, all things on the surface of the earth are related but nearby things are more related than distant things [14], we define a proximity predicate ρ_{prox} to be true when two visual words are within r pixels of each other. That is, given visual words c_i and c_j at locations (x_i, y_i) and (x_j, y_j),

$$c_i \rho_{prox} c_j = \begin{cases} T, & \text{if } \sqrt{(x_i-x_j)^2+(y_i-y_j)^2} \leq r; \\ F, & \text{otherwise.} \end{cases} \tag{3}$$

Thus, the VWCM corresponding to ρ_{prox} indicates the number of times pairs of codewords appear within r pixels of each other in a given image or region. Figure 3(a) shows an example of where ρ_{prox} evaluates to F for two words.

Orientation The SIFT detector provides the orientation of the interest points used to derive the visual words. We postulate that these orientations are indicative of the local shape of image regions and thus derive orientation predicates ρ_{orien} which consider the relative orientations of pairs of visual words.

Given visual words c_i and c_j with (absolute) orientations θ_i and θ_j with respect to some canonical direction such as the x-axis, we define a pair of orientation predicates, one which evaluates to true when the visual words are in-phase (pointing in the same direction) and another which evaluates to true when the visual words are out-of-phase (pointing in opposite directions):

$$c_i \rho_{orien2_1} c_j = \begin{cases} T, & \text{if } cos(\theta_i - \theta_j) \geq 0; \\ F, & \text{otherwise} \end{cases}$$

and

$$c_i \rho_{orien2_2} c_j = \begin{cases} T, & \text{if } cos(\theta_i - \theta_j) < 0; \\ F, & \text{otherwise} \end{cases}$$

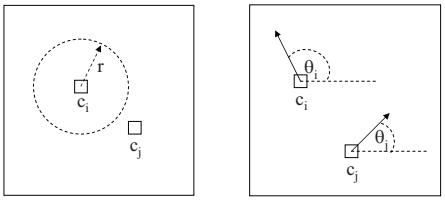
where $-\pi < \theta_i, \theta_j \leq \pi$. Figure 3(b) shows an example of where ρ_{orien2_1} evaluates to T and ρ_{orien2_2} evaluates to F for two words.

We also define a set of four orientation predicates $\rho_{orien4_{1,\ldots,4}}$ which partition the phase space into four bins. That is, the four predicates separately evaluate to true for $\{\sqrt{2}/2 \leq cos(\theta_i - \theta_j)\}$, $\{0 \leq cos(\theta_i - \theta_j) < \sqrt{2}/2\}$, $\{-\sqrt{2}/2 \leq cos(\theta_i - \theta_j) < 0\}$, and $\{cos(\theta_i - \theta_j) < -\sqrt{2}/2\}$.

We characterize the relative instead of absolute orientation of pairs of visual words since overhead imagery generally does not have an absolute reference frame.

4.2. Land-Use Dataset

We evaluate SPCK using a ground truth image dataset of 21 land-use classes. This dataset was manually extracted from aerial orthoimagery downloaded from the United States Geological Survey (USGS) National Map. The images have a resolution of one foot per pixel. 100 images measuring 256×256 pixels were manually selected for each



(a) (b)

Figure 3. We consider spatial predicates which characterize (a) the distance between pairs of visual words, and (b) the relative orientation of pairs of visual words.

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. This particular set of classes was selected because it contains a variety of spatial patterns.

To the best of our knowledge, this is one of the first ground truth datasets derived from *publicly available* high-resolution overhead imagery. This allows us to make it available to other researchers.

4.3. Land-Use Dataset Experiments

We construct visual dictionaries of varying size by applying k-means clustering to over a million SIFT features randomly sampled from images disjoint from the ground truth images. These dictionaries are then used to label SIFT features extracted from the 2,100 ground truth images.

We use an SVM classification framework to compare the different representations and their associated kernels. 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. The only parameter that needs to be estimated is the penalty parameter of the error term. Five-fold cross-validation is performed in which the ground truth dataset is randomly split into five equal sized 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. The results presented below are the average rates over all 21 classes. The SVMs are implemented using the LIBSVM package [1].

We compare the following approaches:

- The "baseline" non-spatial BOVW kernel (sec. 3.2).
- The spatial pyramid match kernel (SPMK) [7] (sec. 3.3).
- The proposed spatial pyramid co-occurrence kernel (SPCK) (sec. 3.6).

Table 1. Classification rates for the land-use dataset. See text for details.

BOVW	SPMK [7]	SPCK	SPCK+	SPCK++
71.86	74.00	73.14	76.05	77.38

- The extended SPCK+ and SPCK++ representations (sec. 3.7).

We also compare the following configurations:

- A proximity predicate alone. This is referred to as SP1 below. We consider distances of r = 20, 50, 100, and 150 pixels.
- A proximity predicate combined with orientation predicates corresponding to a two-bin phase space. This is referred to as SP2 below.
- A proximity predicate combined with orientation predicates corresponding to a four-bin phase space. This is referred to as SP3 below.
- Visual dictionary sizes of 10, 50, and 100 for the co-occurrence component of the SPCK.
- Different numbers of pyramid levels in the SPCK.

4.4. Land-Use Dataset Results

Table 1 compares the best classification rates of the different approaches for the land-use dataset. A visual dictionary size of 100 is used for the BOVW and SPMK approaches as well as the BOVW and SPMK extensions to SPCK. Visual dictionary sizes of 100, 10, and 50 are used for the co-occurrence components of SPCK, SPCK+, and SPCK++. Combined proximity plus 4-bin orientation predicates (SP3) are used for SPCK, SPCK+, and SPCK++.

The land-use dataset is challenging and so the improvement that the proposed SPCK+ and SPCK++ provide over SPMK is significant. In particular, *SPCK++ improves performance over SPMK by more than what SPMK itself improves over the non-spatial BOVW*. And, SPCK+ provides about the same improvement. Note that since SMPK includes BOVW by construction, SPCK+ is a suitable comparison since it is simply SPMK combined with BOVW.

We pick a relatively small BOVW and SPMK dictionary size of 100 for the sake of comparison. SPCK+ and SPCK++ provide a similar improvement over BOVW and SPMK for larger dictionary sizes.

The remainder of this section provides further analysis of the proposed SPCK.

Co-occurrence Dictionary Size Table 2 and figure 4 show the effect of co-occurrence dictionary size on SPCK and SPCK+. The significant result here is that *smaller co-occurrence dictionaries become sufficient or even optimal as the SPCK+ spatial predicates become more specialized*. In particular, a co-occurrence dictionary of just 10 codewords provides better SP3 (SPCK+) performance than one with 50 or 100 codewords. This reduces the computational complexity of SPCK+ to be of the same order as SPMK.

Table 2. The effect of co-occurrence dictionary size (rows) on SPCK and SPCK+. Results are shown for the three different spatial predicate configurations (SP1=proximity only, SP2=proximity+2-bin orientation, SP3=proximity+4-bin orientation), and for the baseline SPCK and extended SPCK+.

	SP1		SP2		SP3	
	SPCK	**SPCK+**	**SPCK**	**SPCK+**	**SPCK**	**SPCK+**
10	58.86	72.33	61.48	74.76	66.43	76.05
50	71.57	74.43	70.90	74.90	73.00	76.00
100	72.62	72.86	72.57	73.14	73.14	74.05

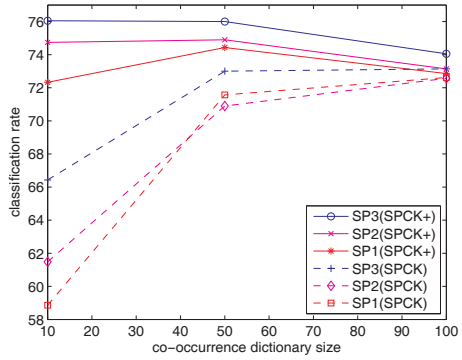


Figure 4. The effect of co-occurrence dictionary size on SPCK and SPCK+.

Table 3. The effect of the spatial predicate proximity distance (rows) on SPCK. Results are shown for different spatial predicate configurations and different co-occurrence dictionary sizes (columns).

	SP1			SP2			SP3		
	10	**50**	**100**	**10**	**50**	**100**	**10**	**50**	**100**
20	58.00	66.52	67.19	60.67	66.38	66.38	62.19	65.71	64.76
50	58.76	69.24	70.81	60.43	69.76	69.81	65.57	70.14	69.14
100	58.62	70.76	72.00	61.38	70.48	72.38	66.29	72.62	72.43
150	58.86	71.57	72.62	61.48	70.90	72.57	66.43	73.00	73.14

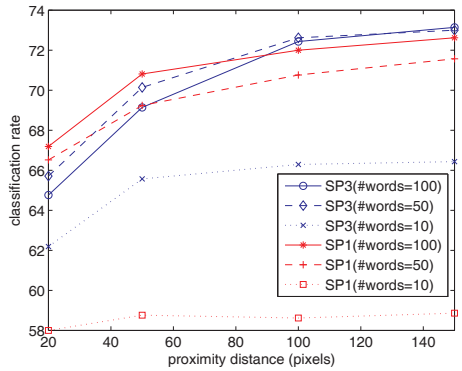


Figure 5. The effect of the spatial predicate proximity distance on SPCK. Results are shown for two different spatial predicate configurations as well as for different co-occurrence dictionary sizes.

Proximity Distance Table 3 and figure 5 show the effect of the spatial predicate proximity distance (r in Eq. 3) on SPCK. Results are shown for the three different spatial predicate configurations as well as for different co-occurrence dictionary sizes. The clear trend is that larger distances improve performance. This indicates that even long range spatial interactions between visual words is important for characterizing the land-use classes.

Table 4. The effect of the number of pyramid levels on SPCK. The rows indicate just level 0, just level 1, just level 2, and all three levels combined. The columns indicate different spatial predicate configurations and co-occurrence dictionary sizes.

	SP1			SP2			SP3		
	10	**50**	**100**	**10**	**50**	**100**	**10**	**50**	**100**
0	52.05	69.81	72.52	55.10	70.19	72.52	60.57	72.57	74.19
1	51.81	66.05	68.00	52.86	65.86	67.57	58.23	67.52	68.33
2	55.01	66.05	66.00	58.52	63.52	62.00	61.19	62.48	59.00
0+1+2	58.86	71.57	72.62	61.48	70.90	72.57	66.43	73.00	73.14

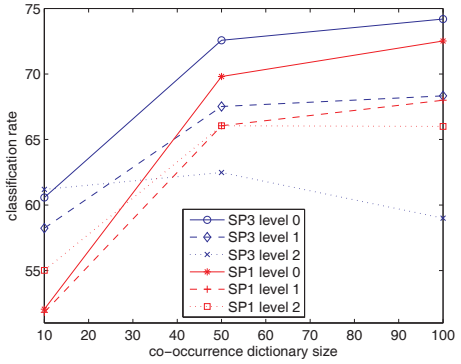


Figure 6. The performance of the individual pyramid levels on SPCK. Results are shown for two different spatial predicate configurations as well as for different co-occurrence dictionary sizes.

Pyramid Levels Table 4 and figure 6 shows the effect of the number of pyramid levels on SPCK. Results are shown for just level 0, just level 1, just level 2, and for all three levels combined. While combining all three levels usually performs best, the interesting trend is that the order of the individual levels depends on the size of the co-occurrence dictionary. In particular, level 0 performs best for a co-occurrence dictionary of size 100 while level 2 performs best for a dictionary of size 10. We will investigate this further in future work.

4.5. Graz-01 Dataset

We also apply our approach to the publicly available dataset Graz-01 [11]. This dataset contains 373 images of category bike, 460 images of category person, and 270 background images as category "counter-class". All the images measure 640x480 pixels and the objects come in different scales, poses, and orientations. This dataset is challenging due to high intra-class variation and have been broadly used as an evaluation dataset in the computer vision community. We evaluate our approach using the same experimental set up as in [11]. In particular, our training set contains 100 positive images (bike or person) and 100 negative images from the other two categories, where half are from the background and half are from the other object category. Our test set consists of 100 images with a similar distribution to the training set. We report equal error rates averaged

Table 5. Evaluation using the Graz-01 dataset. Comparison of the proposed SPCK+ approach with Boosting+SIFT [11], SPMK [7], PDK [8], and NBNN [2].

	[11]	SPMK [7]	PDK [8]	NBNN [2]	SPCK+
Bike	86.5	86.3±2.5	90.2±2.6	90.0±4.3	91.0±4.8
Person	80.8	82.3±3.1	87.0±3.8	87.0±4.6	87.2±3.8

Table 6. Results on the 15 Scene dataset.

	SPMK [7]	SPCK++
Classification Rate	81.40±0.50	82.51±0.43

over ten runs.

Table 5 compares our technique with other approaches that characterize the spatial arrangement of visual words, namely the Boosting+SIFT approach of Opelt et al. [11], the SPMK approach of Lazebnik et al. [7], the proximity distribution kernel (PDK) approach of Ling and Soatto [8], and the naive Bayes nearest neighbor (NBNN) approach of Boiman et al. [2] (while NBNN is not a spatial based approach, we include it here for completeness). Our SPCK+ is shown to perform better than the other approaches.

4.6. 15 Scene Dataset

Finally, we apply our approach to the publicly available 15 Scene dataset [7]. This dataset contains a total of 4485 images in 15 categories varying from indoor scenes such as store, bedroom, and kitchen, to outdoor scenes such as coast, city, and forest. Each category has between 200 to 400 images and each image measures approximately 300x300 pixels. Following the same experiment setup as [7], we randomly pick 100 images per category for training and use the rest for testing. Table 6 compares our results with those of SPMK. We see again that our approach SPCK++ improves over SPMK.

The images in the 15 Scene dataset tend to be strongly aligned so that local spatial arrangement tends to be less important than global layout. The proposed approach thus results in only a modest improvement over SPMK (and does not beat the best published results) on this dataset since it is designed to distinguish between image classes that possibly differ only in their relative spatial arrangements such as the land-use dataset above. The global alignment of the 15 Scene dataset is a much stronger signal for discriminating between classes than relative spatial arrangement. It is for these same reasons that SPCK is not appropriate for strongly aligned object class datasets such as Caltech-101.

5. Conclusion

We proposed spatial pyramid co-occurrence, a novel approach to characterizing the photometric and geometric aspects of an image. The representation captures both the absolute and relative spatial arrangements of visual words and can characterize a wide variety of spatial relationships through the choice of the underlying spatial predicates.

We performed a thorough evaluation using a challenging 21 land-use class dataset which can be made publicly available since it was derived from royalty free imagery. The proposed approach was shown to perform better on this dataset than a non-spatial bag-of-visual-words approach as well as a popular approach for characterizing the absolute spatial arrangement of visual words. And, while our primary objective is analyzing overhead imagery, we also demonstrated that our approach achieves state-of-the-art performance on the Graz-01 object class dataset and performs competitively on the 15 Scene dataset.

We noted several salient aspects of our approach. In particular, we demonstrated that small visual dictionaries become optimal as the spatial predicates become more specialized. This tradeoff is an interesting result which we will investigate further in future work.

6. 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.

References

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

[2] O. Boiman, E. Shechtman, and M. Irani. In defense of nearest-neighbor based image classification. In *CVPR*, 2008.

[3] G. Csurka, C. R. Dance, L. Fan, J. Willamowski, and C. Bray. Visual categorization with bags of keypoints. In *In Workshop on Statistical Learning in Computer Vision, ECCV*, 2004.

[4] K. Grauman and T. Darrell. The pyramid match kernel: Discriminative classification with sets of image features. In *ICCV*, 2005.

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

[6] J. Huang, S. R. Kumar, M. Mitra, W.-J. Zhu, and R. Zabih. Image indexing using color correlograms. In *CVPR*, 1997.

[7] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR*, 2006.

[8] H. Ling and S. Soatto. Proximity distribution kernels for geometric context in category recognition. In *ICCV*, 2007.

[9] D. Liu, G. Hua, P. Viola, and T. Chen. Integrated feature selection and higher-order spatial feature extraction for object categorization. In *CVPR*, 2008.

[10] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004.

[11] A. Opelt, M. Fussenegger, A. Pinz, and P. Auer. Weak hypotheses and boosting for generic object detection and recognition. In *ECCV*, 2004.

[12] S. Savarese, J. Winn, and A. Criminisi. Discriminative object class models of appearance and shape by correlatons. In *CVPR*, 2006.

[13] J. Sivic and A. Zisserman. Video Google: A text retrieval approach to object matching in videos. In *ICCV*, 2003.

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