Geographic Image Retrieval Using Local Invariant Features

Yi Yang, *Student Member, IEEE*, and Shawn Newsam, *Member, IEEE*

***Abstract*—This paper investigates local invariant features for geographic (overhead) image retrieval. Local features are particularly well suited for the newer generations of aerial and satellite imagery whose increased spatial resolution, often just tens of centimeters per pixel, allows a greater range of objects and spatial patterns to be recognized than ever before. Local invariant features have been successfully applied to a broad range of computer vision problems and, as such, are receiving increased attention from the remote sensing community particularly for challenging tasks such as detection and classification. We perform an extensive evaluation of local invariant features for image retrieval of land-use/land-cover (LULC) classes in high-resolution aerial imagery. We report on the effects of a number of design parameters on a bag-of-visual-words (BOVW) representation including saliency- versus grid-based local feature extraction, the size of the visual codebook, the clustering algorithm used to create the codebook, and the dissimilarity measure used to compare the BOVW representations. We also perform comparisons with standard features such as color and texture. The performance is quantitatively evaluated using a first-of-its-kind LULC ground truth data set which will be made publicly available to other researchers. In addition to reporting on the effects of the core design parameters, we also describe interesting findings such as the performance-efficiency tradeoffs that are possible through the appropriate pairings of different-sized codebooks and dissimilarity measures. While the focus is on image retrieval, we expect our insights to be informative for other applications such as detection and classification.**

***Index Terms*—Bag of visual words, content-based image retrieval, high-resolution overhead image analysis, land cover, land use, local invariant features, remote sensing.**

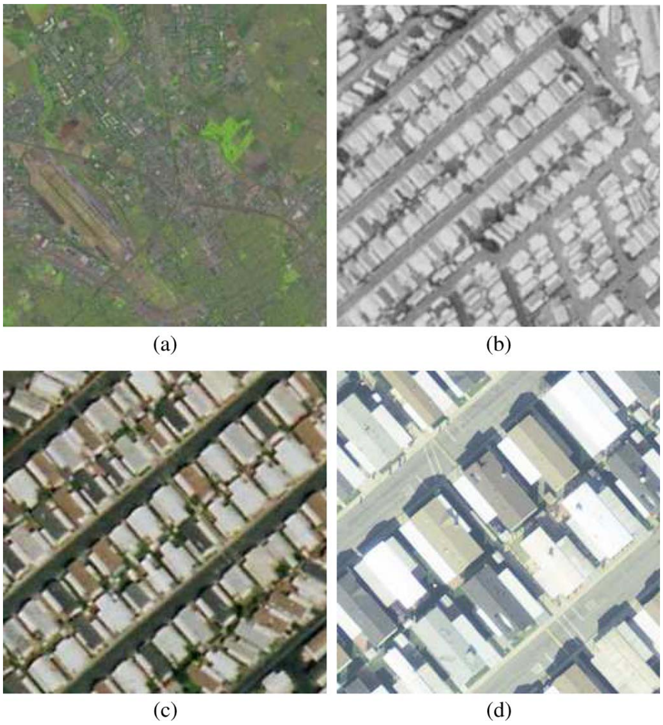

(a) (b) (c) (d)

Fig. 1. Images with resolutions of (a) 30 m, (b) 1 m, (c) 2 ft (approximately 60 cm), (d) and 1 ft (approximately 30 cm). The increased resolution of newer imagery supports analysis methods that were not possible before such as approaches based on local features which characterize individual objects and their components instead of patterns.

I. Introduction

The increased spatial resolution and coverage of overhead imagery from satellites and aircraft provide novel opportunities for advancing the field of remote sensed image analysis, particularly with regard to automated image understanding. A greater range of objects and spatial patterns can be observed than ever before due to the increased resolution. Fig. 1(a)–(d) shows images with spatial resolutions of 30 m, 1 m, 2 ft (approximately 60 cm), and 1 ft (approximately 30 cm). The image in Fig. 1(a) is from Landsat V which was launched in 1999. The images in Fig. 1(b) and (c) are aerial images with approximately the same resolutions as IKONOS which was launched in 2000 and Quickbird which was launched in 2001. Imagery with the resolution of the aerial image in Fig. 1(d) or even higher is now available for large geographic regions.

The increased resolution of the newer imagery supports analysis methods which were not possible before. This paper investigates one such class of methods wherein local image regions are *characterized* by features designed to be invariant to differences in appearance resulting from geometric transformations such as rotation or scaling as well as from photometric transformations such as changes in illumination. The image regions themselves are also *detected* in an invariant manner. These so-called local invariant features have been successfully applied to a range of standard (nongeographic) computer vision problems, and there has been increasing interest in using them for overhead image analysis.

The fundamental contribution of this paper is an investigation into local invariant features for overhead image retrieval. To our knowledge, it is the first study of its kind. We perform an

Manuscript received June 14, 2011; revised December 22, 2011 and April 22, 2012; accepted May 6, 2012. Date of publication August 2, 2012; date of current version January 17, 2013. This work was supported in part by the U.S. Department of Energy under an Early Career Scientist and Engineer/PECASE Award and by the U.S. National Science Foundation under Grant IIS-0917069.

The authors are with the Electrical Engineering and Computer Science Program, University of California, Merced, CA 95343 USA (e-mail: yyang6@ucmerced.edu; snewsam@ucmerced.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TGRS.2012.2205158

extensive evaluation of local invariant features for image retrieval of land-use/land-cover (LULC) classes in high-resolution aerial imagery. We report on the effects of a number of design parameters on a bag-of-visual-words (BOVW) representation including saliency- versus grid-based local feature extraction, the size of the visual codebook, the clustering algorithm used to create the codebook, and the dissimilarity measure used to compare the BOVW representations. We also perform a comparison with standard features such as color and texture. The performance is quantitatively measured using a first-of-its-kind LULC ground truth data set which will be made publicly available to other researchers. While the focus is on image retrieval, we expect our insights to be informative for other applications of local invariant features such as detection and classification.

II. Background

The remote sensing community has begun to realize the potential for local feature-based analysis of high-resolution imagery. A number of methods have been developed to perform image matching for registration [1]–[7] and change detection [8], [9]. Closer to the work presented in this paper, researchers have also investigated local features for detection and classification. Sirmacek and Unsalan [10]–[12] use local features to detect buildings and urban areas in 1-m resolution IKONOS imagery. Xu *et al.* [13] compare quantized color and texture features with local features for classifying 0.25-m resolution aerial image regions into four LULC classes. Chen *et al.* [14] also compare local features with standard color and texture features to classify 0.5-m Digital Globe imagery into 19 LULC classes. Skurikhin [15] investigates attention-based saliency detection to perform local feature-based classification of 0.5-m resolution Digital Globe and Google Earth imagery into anthropogenic or natural regions. Gleason *et al.* [16] and Vatsavai *et al.* [17] use quantized local features to detect complex geospatial objects such as nuclear and coal power plants in 1-m resolution Digital Globe imagery. Ozdemir and Aksoy [18] investigate graph-based spatial arrangements of quantized local features to classify 1-m resolution IKONOS imagery into eight LULC classes. Bordes and Prinet [19] investigate spatial correlograms of quantized local features to classify high-resolution Digital Globe imagery into eight LULC classes.

Our work on local invariant features differs from that above in three fundamental ways. First, we perform a thorough investigation into a range of design parameters such as the size of the visual dictionary used to quantize the local features, whether k-means clustering should be applied using the Euclidean or Mahalanobis distance when constructing the dictionary, and the relative performance of nine different dissimilarity measures for comparing histograms of quantized local features. We feel such an investigation is timely due to the increased application of local invariant features for overhead image analysis. (Indeed, ours is the most thorough investigation of these parameters for any image analysis problem, not just overhead imagery.) Second, we consider 21 different LULC classes, significantly more than any of the above works except that of Chen *et al.* [14] who consider 19. Third, we perform our evaluations on high-resolution aerial imagery *that is in the public domain*. We construct a ground truth data set of the 21 classes using imagery acquired from the U.S. Geological Survey (USGS) National Map. We will make this data set publicly available to other researchers as a standard for comparing methods (and thus this paper also establishes benchmarks on this data set for other researchers to improve upon). Such standardized data sets have proven critical for developing improved image classification techniques in non-overhead imagery (e.g., the Caltech 101 [20] and 256 [21] object class image data sets, and the PASCAL visual object classes data sets [22]). To our knowledge, ours is the first publicly available high-resolution LULC evaluation data set.

Content-based image retrieval (CBIR) has been an active area of research in computer vision since the mid-1990s. Motivated by the need to provide effective access to the growing collections of digital images, systems starting with IBM's Query by Image Content from 1995 [23] have been proposed which automatically annotate images using visual features such as color, texture, and shape. While the shortcomings of these low-level features to capture high-level, semantic concepts has been well-documented, CBIR remains *an effective framework in which to investigate visual features particularly for computing image similarity*. Pair-wise image comparison is fundamental to a range of kernel-based machine learning techniques such as support vector machines and so investigations such as ours stand to inform applications beyond retrieval.

A number of works have investigated different features for performing image retrieval in large collections of geographic images. Similar to other domains, researchers have investigated intensity features [24], spectral (color) features [25], [26], shape features [27]–[30], structural features [31], [32], texture features [29], [30], [33]–[37], and combinations thereof such as multi-spectral texture features [38]. However, to our knowledge, ours is the first work to investigate local invariant features for geographic image retrieval.

To summarize, the novel contributions of this paper include:

- The first study of local invariant features for content-based geographic image retrieval, in particular showing their superiority over standard features such as color and texture.
- The most thorough investigation of the effects of different design parameters of local invariant features for any image analysis problem, not just overhead imagery.
- A first-of-its-kind 21 LULC data set which will be made publicly available to other researchers. We anticipate this will serve as a standardized data set for comparing techniques, something which has greatly helped other applications of image analysis but has largely been lacking in the remote sensing field.

III. Local Invariant Features

A. Desirable Properties

There are generally two steps to using local invariant features for image analysis. First, a *detection* step identifies interesting locations in the image usually according to some measure of saliency. These are termed interest points. Second, a *descriptor* is computed for each of the image patches centered at the interest points. The following describes the properties of the detection and descriptor that contribute to the effectiveness of local invariant features.

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. Note that occlusion includes the case where part of an object is hidden as well as the case where the object is cropped by the edge of the image.

Invariance: Local image analysis has a long history including corner and edge detection [39]. 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 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 for example) 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 [40], we choose a detector that is invariant up to similarity transformations only 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 [40]. Invariance to translation and scale is typically accomplished through scale-space analysis with automatic scale selection [41]. Invariance to rotation is typically accomplished by estimating the dominant orientation of the gradient of a scale-normalized image patch. We construct the evaluation data set in the experiments below to contain regions and objects that occur at arbitrary and varying orientations as is generally the case in overhead imagery.

Photometric image transformations result from variations in illumination intensity and direction. Photometric invariance is typically obtained in both the detection and descriptor by simply modeling 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. We construct the data set used in the experiments below to contain images acquired under a range of different illumination conditions and from a number of different optical sensors.

Robust yet Distinctive: The features should be robust to other image transformations for which they are not designed to be invariant through explicit modeling. 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 [42] has shown that local invariant features achieve this balance. The evaluation data set below contains images of varying quality.

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 natural images present representation and computational challenges. The histograms of quantized descriptors used in this work have shown to be an effective and efficient method to help mitigate the associated costs.

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 using local features has been demonstrated in prototype systems [43] as well as in commercial products such as the SnapTell camera-phone recognition application [44].

B. SIFT Features

We choose David Lowe's scale invariant feature transform (SIFT) [45], [46] as the interest point detector and descriptor. While there are other detectors, such as the Harris-Laplace/Affine [40], Hessian-Laplace/Affine [40], Kadir and Brady's Saliency Detector [47]; other descriptors, such as shape context [48], steerable filters [49], PCA-SIFT [50], spin images [51], moment invariants [52], and cross-correlation; and other detector/descriptor combinations, such as maximally stable extremal regions [53] and speeded up robust features [54], 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 as described above. Second, an extensive comparison with other local descriptors found that the SIFT descriptor performed the best in an image matching task [42]. We note, however, that the primary contribution of this paper is to demonstrate that local invariant features are effective for geographic image retrieval and to perform a thorough investigation into the BOVW design parameters. We expect that our findings to be largely independent of the underlying detector and descriptor.

As mentioned above, SIFT descriptors are extracted from an image in two steps. 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 characterized under different image acquisition conditions.

1) 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

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

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 3-D DoG scale space is compared with its eight spatial neighbors at the same scale, and with its 18 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 (RI).

2) 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. (For our evaluation data set below, most patches range in diameter from 6 to 50 pixels with a few that are larger.) This makes the descriptor scale invariant. The feature descriptor consists of histograms of gradient directions computed over a 4 × 4 spatial grid. The interest point orientation estimate is used to align the gradient directions to make the descriptor RI. The gradient directions are quantized into eight bins so the final feature vector has dimension 128 $(4 \times 4 \times 8)$. 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.

We also consider extracting SIFT descriptors from a fixed grid instead of from the salient interest points. We refer to this as *grid-based* feature extraction. It is often all called dense sampling as it typically results in a larger number of descriptors since interest points are not detected in non-salient regions (of uniform intensity for example). We refer to the standard approach as *saliency-based* feature extraction.

C. 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 descriptors have 128 dimensions. We adopt a standard approach, termed BOVW [55], to summarize the descriptors without regard to where they appear in an image. The analogy to representing text documents as word count frequencies is made possible by quantizing the 128 dimension SIFT descriptors. We apply standard k-means clustering to a large number of SIFT descriptors to create a dictionary of visual words or codebook. Descriptors extracted from novel images are then quantized by assigning the label of the closest cluster centroid or codeword. The final representation is the frequency or histogram of the codewords in an image

$$h_{INT} = [t_0, t_1, \ldots, t_{k-1}]$$

where t_i is number of times codeword i appears. In the experiments below, we investigate a number of different BOVW design parameters such as the size of the visual dictionary (k in the k-means clustering), the number of SIFT descriptors to which the clustering is applied, and whether the Mahalanobis or Euclidean measure is used to compute the descriptor-to-cluster-centroid distance for point (re)assignment during the iterative k-means algorithm and during quantization.

IV. Data set

We perform quantitative evaluation of retrieval performance using a manually constructed ground truth data set. The data set consists of images of 21 LULC classes selected from aerial orthoimagery with a pixel resolution of 30 cm. Large images were downloaded from the 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 by 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. Five samples of each class are shown in Fig. 2. The images downloaded from the National Map are in the red-green-blue (RGB) color space. Both RGB and grayscale versions of the 2100 ground truth images are used where $Gray = 0.299*R + 0.587*G + 0.114*B$.

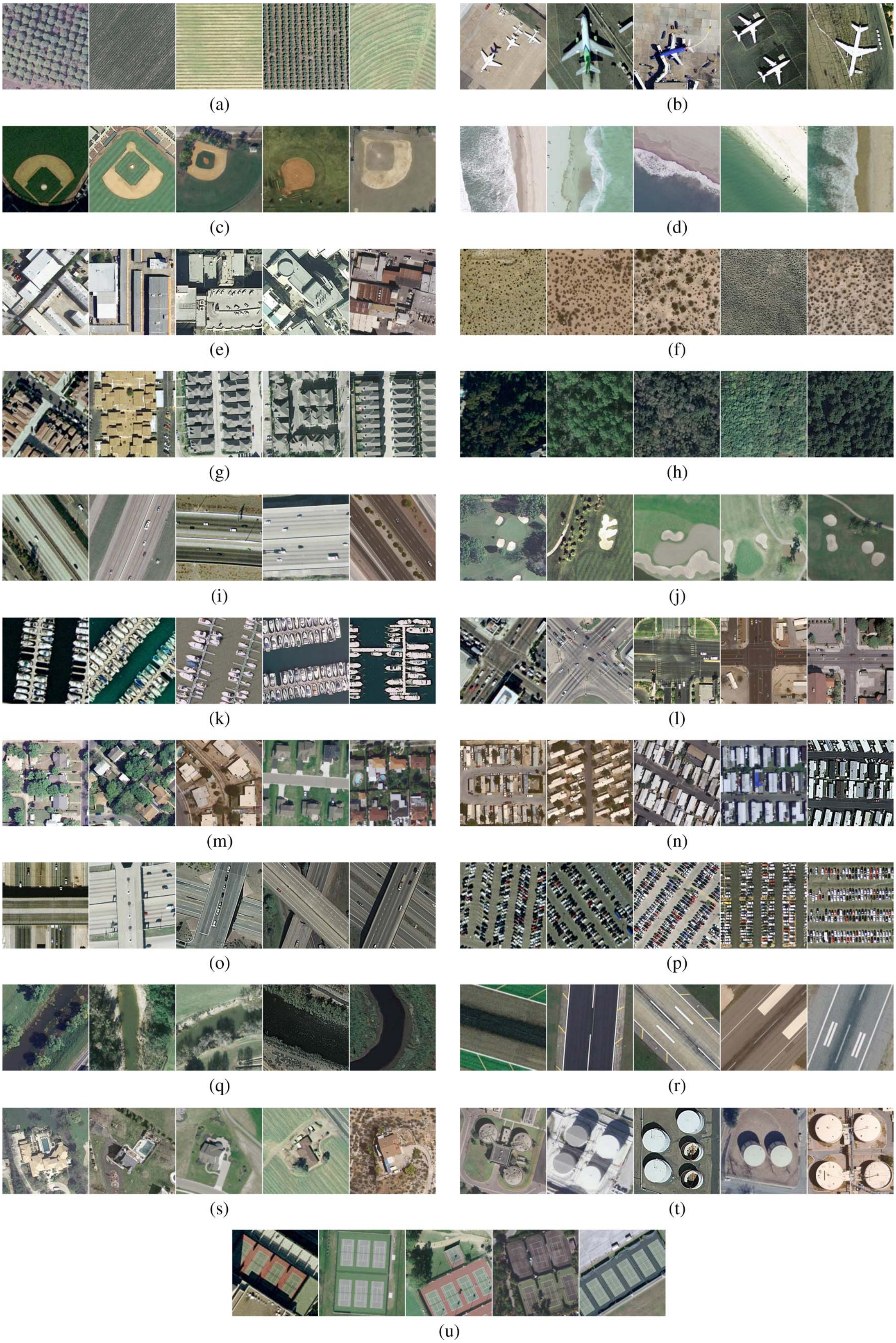
A significant benefit of using aerial orthoimagery from the USGS National Map is that the data is already in the public domain. Thus, our 21 class LULC data set is the largest data set of its kind that can be made publicly available to other researchers. We will make the data set available through our research group's website.

V. Experiments

This section describes the image retrieval protocol used in the experiments. It first describes the standard color and texture features against which the local features are compared. It then describes the different configurations of the BOVW representation used to summarize the local features for an image. The different dissimilarity measures considered are then described. Finally, the retrieval performance metrics used for the quantitative evaluation are described.

A. Standard Features

Three standard image features are considered: simple statistics, homogeneous texture, and color histogram features.



(a) (b) (c) (d) (e) (f) (g) (h) (i) (j) (k) (l) (m) (n) (o) (p) (q) (r) (s) (t) (u)

Fig. 2. Ground truth data set contains 100 images from each of 21 land-use/land-cover classes. Five samples from each class are shown above. The data set will be made publicly available to other researchers. (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.

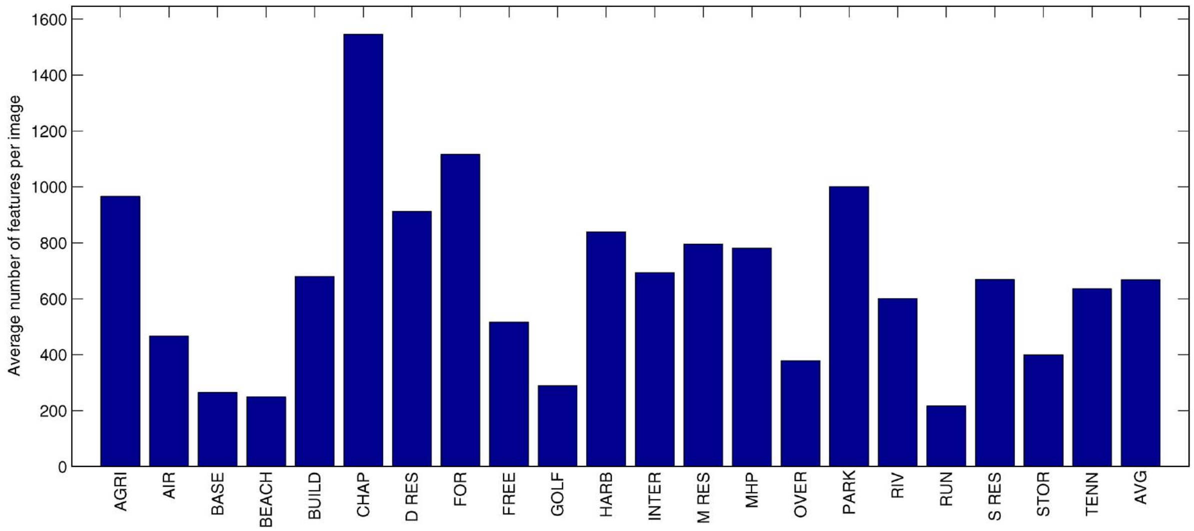


Fig. 3. Average number of features per class for saliency-based local feature extraction.

1) Simple Statistics: A 2-D feature vector is computed for each ground truth image consisting of the mean and standard deviation of the grayscale values:

$$f_{SS} = (\mu, \sigma).$$

This is referred to as the simple statistics feature and serves as a baseline for the experiments.

2) Homogeneous Texture: Homogeneous Texture Descriptors compliant with the MPEG-7 Multimedia Content Description Interface [56] are extracted using banks of Gabor filters tuned to five scales and six orientations. A 60-dimensional feature vector is formed from the mean and standard deviation of the 30 filters

$$f_{texture} = [\mu_{11}, \sigma_{11}, \mu_{12}, \sigma_{12}, \ldots, \mu_{1S}, \sigma_{1S}, \ldots, \mu_{RS}, \sigma_{RS}]$$

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. To account for differences in range, normalized versions of the features are also produced in which each of the $2RS$ components is scaled to have a mean of zero and a standard deviation of one over the ground truth data set.

3) Color Histogram: Color histogram features are computed in three color spaces: RGB, hue lightness saturation (HLS), and CIE Lab. Each dimension is quantized into eight bins for a total histogram feature length of 512. The histograms are normalized to sum to one (L1 norm equal to one). This results in three different color histogram features: f_{RGB}, f_{HLS}, and f_{Lab}.

B. Local Invariant Features

128 dimensional local invariant descriptors are extracted for each ground truth image using the SIFT descriptor algorithm. As described above, we consider two extraction modes, saliency-based extraction using the SIFT detector and grid-based extraction. Saliency-based extraction results in a mean of 668 descriptors for each 256 × 256 image over all classes. The runway class tends to have the fewest descriptors per image with a mean of 218 and the forest class has the most with a mean of 1117. Fig. 3 indicates the per class means.

Grid-based feature extraction is performed using three different grid spacings, 4-pixel, 8-pixel, and 16-pixel, which result in 3721, 961, and 256 features per image, respectively.

The SIFT descriptors are quantized using codebooks resulting from applying k-means clustering to a large number of SIFT descriptors sampled at random from the large aerial images from which we created the evalution data set. The 2100 images in the evaluation data set represent less than four percent of the total area in the large images. Thus, for all practical purposes, there is no overlap between the images used to create the codebooks and the evaluation data set. The codebooks are thus not specific to the particular images in the evaluation data set. Further, the codebooks are not specific to the 21 classes (as they are based on SIFT features randomly sampled from the diverse large aerial images), and we would expect them to generalize to additional classes.

Codebooks are created using k-means for:

- A wide range of different numbers of clusters (k): 10, 25, 50, 75, 100, 125, 150, 175, 200, 250, 300, 350, 400, 450, 500, 600, 700, 800, 900, 1000, 2000, 3000, 4000, 5000, 6000, 7000, 8000, 9000, 10 000, 15 000, 20 000.
- Different-sized sets of randomly sampled points: 100 thousand and one million.
- Different distance measures: Euclidean and Mahalanobis.

Each clustering is performed ten times using a different set of randomly sampled points.

SIFT histogram features are calculated for each ground truth image by using a codebook to quantize the SIFT descriptors extracted from the image. The histogram features thus range in length from 10 to 20 000 components. Three versions of the histogram features are considered: 1) unnormalized SIFT histogram features which simply contain the codeword counts; 2) L1 normalized SIFT histogram features where the components are normalized to sum to one; and 3) L2 normalized SIFT histogram features where the components are normalized so the feature vectors have length one.

C. Dissimilarity Measures

Each image in the ground truth data set is represented by a multidimensional feature vector. This is either a 2-D simple statistics feature, a 60-dimensional texture feature, a 512-dimensional color histogram feature, or a k-dimensional SIFT histogram feature where k is the size of the codebook used to quantize the SIFT descriptors. The dissimilarity measure used to compare two images depends on the type of feature. (While a few of the measures below are technically similarity measures, we refer to them as dissimilarity measures for consistency. A similarity measure can be treated as a dissimilarity measure for retrieval by simply reversing the ranking of the retrieved set.)

1) Simple Statistics: The dissimilarity between two images with simple statistics features $f1$ and $f2$ is computed using the L2 or Euclidean distance

$$d_{SS}(f1, f2) = \|f1 - f2\|_2 = \sqrt{(\mu 1 - \mu 2)^2 + (\sigma 1 - \sigma 2)^2}.$$

2) Texture: The default dissimilarity between two images with texture features $f1$ and $f2$ is also computed using the L2 distance

$$d_{texture}(f1, f2) = \|f1 - f2\|_2 = \sqrt{\sum_{i=1}^{2RS} (h1_i - h2_i)^2}.$$

This results in an orientation (and scale) sensitive dissimilarity measure. Orientation invariant similarity is possible by using a modified distance function

$$d_{RI}(f1, f2) = \min_{r \in R} \left\| f1_{\langle r \rangle} - f2 \right\|_2$$

where $f_{\langle r \rangle}$ represents f circularly shifted by r orientations

$$\begin{aligned} f_{\langle r \rangle} = \big[&(f_{r1}, f_{r2}, \ldots, f_{rS}), \left(f_{(r+1)1}, f_{(r+1)2}, \ldots, f_{(r+1)S}\right), \\ &\ldots, (f_{R1}, f_{R2}, \ldots, f_{RS}), (f_{11}, f_{12}, \ldots, f_{1S}), \ldots, \\ &\left(f_{(r-1)1}, f_{(r-1)2}, \ldots, f_{(r-1)S}\right) \big]. \end{aligned}$$

Note that parentheses have been added for clarity. We refer to this as the RI texture distance measure. Conceptually, this distance function computes the best match between rotated versions of the images without repeating the feature extraction. The granularity of the rotations is of course limited by the filter bank construction.

3) Color Histogram: A number of different histogram distance measures are used to compute the dissimilarity between pairs of images with respect to color histogram features, including: Bhattacharyya, chi-square, correlation, cosine, inner product, intersection, L1, L2, and Earth Mover's Distance (EMD). For two images with color histogram features $f1$ and $f2$ of dimension d, the first eight of these are as follows:

$$d_{Bhattacharyya} = \sqrt{1 - \sum_{i=1}^{d} \frac{\sqrt{f1_i f2_i}}{\sqrt{\sum_{j=1}^{d} f1_j \sum_{k=1}^{d} f2_k}}}$$

$$d_{chi-square} = \sum_{i=1}^{d} \frac{(f1_i - f2_i)^2}{f1_i + f2_i}$$

$$d_{correlation} = \frac{\sum_{i=1}^{d} f1'_i f2'_i}{\sqrt{\sum_{j=1}^{d} f1'_j f2'_j}}$$

$$\text{where } f' = f - \frac{1}{d}\sum_{i=1}^{d} f$$

$$d_{cosine} = \frac{\sum_{i=1}^{d} f1_i f2_i}{\sqrt{\sum_{j=1}^{d} f1_j^2 \sum_{k=1}^{d} f2_k^2}}$$

$$d_{innerproduct} = \sum_{i=1}^{d} f1_i f2_i$$

$$d_{intersection} = \sum_{i=1}^{d} \min(f1_i, f2_i)$$

$$d_{L1} = \|f1 - f2\|_1 = \sum_{i=1}^{d} \|f1_i - f2_i\|$$

$$d_{L2} = \|f1 - f2\|_2 = \sqrt{\sum_{i=1}^{d} (f1_i - f2_i)^2}.$$

The EMD [57] measures the distance between two distributions, in our case histograms, by viewing the distributions as "piles of dirt" and computing the cost of turning one pile into another. The cost is the amount of dirt times the distance it is moved. We consider two cases. One, the default, in which the distance between histogram bins is simply the Euclidean distance between the bin indices (the color histograms are 3-D). In the other case, a cost matrix indicates the actual distance between histogram bins in color space.

4) SIFT Histogram: The same histogram distance measures are used to compare the SIFT histogram features. Only the cost matrix version of the EMD is used as the bin indices of the SIFT histogram features provide no information on the relations between the bins. The cost matrix is computed as the Euclidean distances between the 128-dimensional centroids corresponding to the bins.

D. Retrieval Performance

The features and associated dissimilarity measures are used to perform image retrieval as follows. Let T be a collection of M images; let f^m be the feature vector extracted from image m, where $m \in 1, \ldots, M$; let $d(\cdot, \cdot)$ be a distance function defined on the feature space; and let f^{query} be the feature vector corresponding to a given query image. Then, the image in T

most similar to the query image is the one whose feature vector minimizes the distance to the query's feature vector

$$m^* = \underset{1 \le m \le M}{\arg\min}\, d(f^{query}, f^m).$$

Likewise, the k most similar images are those that result in the k smallest distances when compared to the query image. Retrieving the k most similar items is commonly referred to as a k-nearest neighbor (kNN) query. (Note that this k is distinct from the k used in the k-means clustering for creating the codebooks.)

Given a ground-truth data set, there are a number of ways to evaluate retrieval performance. A single measure of performance that considers both the number and order of the ground truth items that appear in the top retrievals is the average normalized modified retrieval rank (ANMRR) which was used extensively in the MPEG-7 standardization process [56]. Consider a query q with a ground-truth size of $NG(q)$. The $Rank(k)$ of the kth ground-truth item is defined as the position at which it is retrieved. A number $K(q) \geq NG(q)$ is chosen so that items with a higher rank are given a constant penalty

$$Rank(k) = \begin{cases} Rank(k), & \text{if } Rank(k) \leq K(q) \\ 1.25K(q), & \text{if } Rank(k) > K(q). \end{cases}$$

$K(q)$ is commonly chosen to be $2NG(q)$. The *average rank* (AVR) for a single query q is then computed as

$$AVR(q) = \frac{1}{NG(q)} \sum_{k=1}^{NG(k)} Rank(k).$$

To eliminate influences of different $NG(q)$, the NMRR

$$NMRR(q) = \frac{AVR(q) - 0.5\,[1 + NG(q)]}{1.25K(q) - 0.5\,[1 + NG(q)]}$$

is computed. $NMRR(q)$ takes values between zero (indicating whole ground truth found) and one (indicating nothing found) irrespective of the size of the ground-truth for query q, $NG(q)$. Finally, the ANMRR can be computed for a set NQ of queries

$$ANMRR = \frac{1}{NQ} \sum_{q=1}^{NQ} NMRR(q).$$

ANMRR ranges in value between zero to one with lower values indicating better retrieval performance.

VI. Results

Table I summarizes the best results for the different features considered in this study in terms of ANMRR values averaged over all 21 classes. The local invariant features are shown to perform better than the standard features. The best performance for the texture features results from using the RI measure to compare unnormalized features. The best performance for the color descriptors results from using the EMD cost matrix measure to compare HLS histograms. The best performance for the local descriptors results from using the L1 measure to compare L1 normalized histograms based on a codebook of 15 000 words created using k-means clustering with the Euclidean distance and descriptors extracted using the saliency-based method. Fig. 4 shows the precision-recall curves corresponding to these configurations. (Precision is the fraction of correct retrievals and recall is the fraction of ground truth items retrieved for a given result set.) Precision and recall are calculated as the number of retrievals is varied from 1 to 2100, and the plots are the average taken over all 2100 queries. The saliency-based local features result in higher precision for all but the lowest recall levels.

TABLE I
Summary of Best Results as Measured Using ANMRR Values Averaged Over the 21 Classes. See Section VI for the Optimal Configurations That Produced These Results

Feature	ANMRR	Time (sec.)
Simple Statistics	0.8079	0.3300
Texture	0.6304	40.04
Color Histogram	0.7351	1.654E+05
Local Features	*0.5914*	193.3

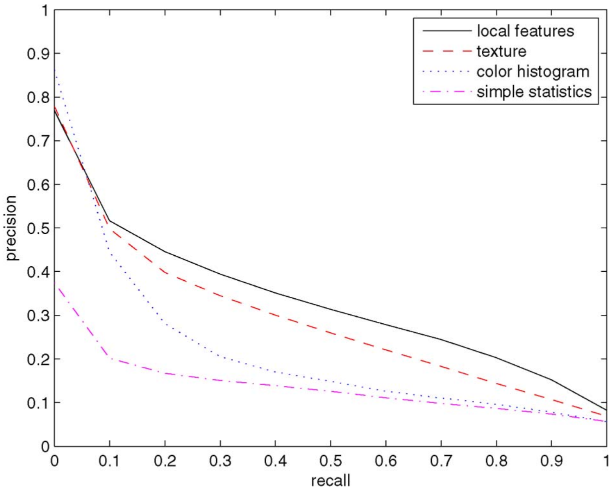


Fig. 4. Precision-recall curves for the different features. Saliency-based local invariant features result in higher precision for all but the lowest recall levels.

TABLE II
Performance of Texture Features

Features	Dissimilarity Measure	ANMRR	Time (sec.)
Unnormalized	Orientation Selective	0.6957	0.8500
Normalized	Orientation Selective	0.7036	0.8600
Unnormalized	Rotation Invariant	*0.6304*	40.04
Normalized	Rotation Invariant	0.6555	40.47

Table I also indicates how long the similarity retrievals took in seconds as an empirical comparison of the computational complexity of the different descriptors and their distance measures. The table shows the number of seconds required to perform 2100 queries in which the pairwise distance between a query and 2100 target images is computed and then used to order the result sets. These are approximate timings on a standard desktop machine and are provided for comparison purposes.

The remainder of this section describes the performance of the specific features in more detail.

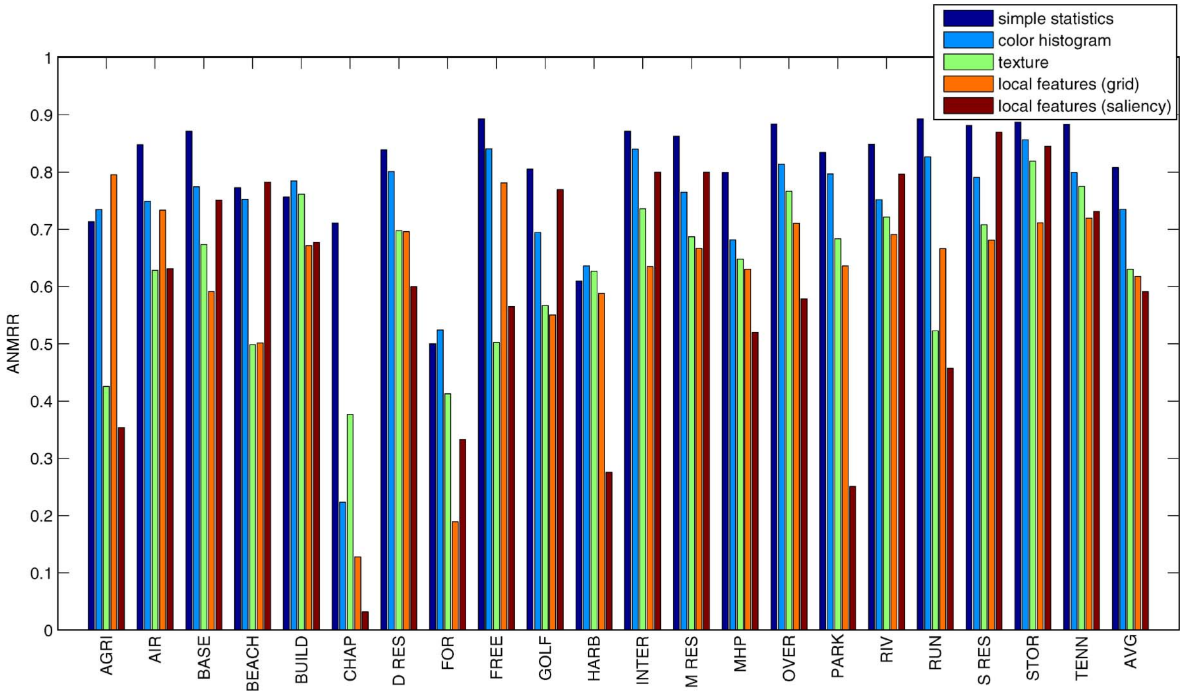
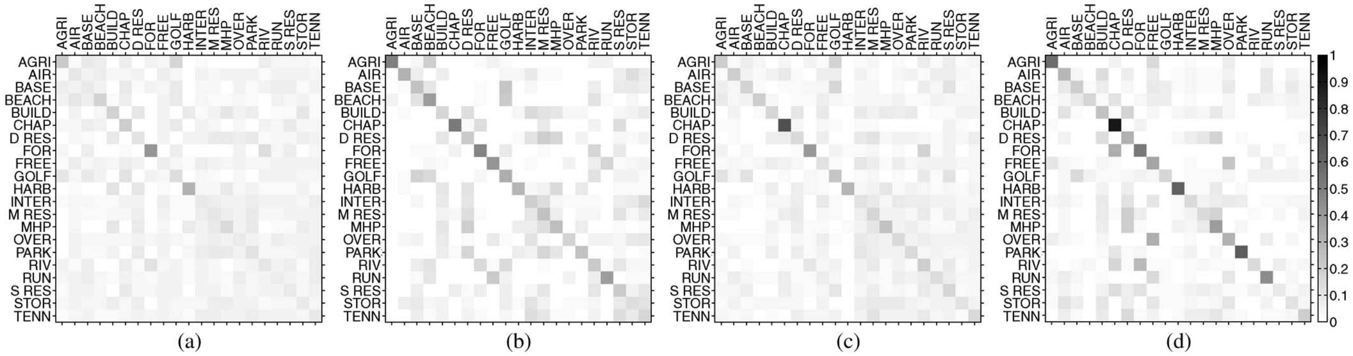


Fig. 5. Per class performance corresponding to the optimal feature configurations.



(a) (b) (c) (d)

Fig. 6. Confusion matrices corresponding to (a) simple statistics features, (b) texture features, (c) color histogram features, and (d) saliency-based local features. The rows indicate the query class and the columns the target classes. For each query image, we record the fraction of images in the top 100 retrievals that are in each of the 21 target classes. These values are then averaged over all 100 query images.

A. *Texture*

Table II shows the ANMRR values and timings of different texture feature configurations. The RI distance measure performs better than the orientation selective one. This makes sense because the image classes do not have a preferred orientation: either they do not have a distinct orientation—e.g., chaparral—or, if they do, it is not consistent—e.g., beach. The RI measure does take considerably longer as expected due to its increased computational complexity.

The unnormalized texture features perform better than the L2 normalized ones. This is true for both distance measures. This is an interesting result which, to the best of our knowledge, has not been investigated or reported before. Previous applications of texture features based on Gabor filters [58] usually perform L2 normalization to account for different dynamic ranges between the feature components. However, the design of the filterbanks [35] includes scaling factors to compensate for the different regions of support, and so our results indicate that further normalization suppresses discriminative frequency information and results in decreased performance.

Fig. 5 compares the per-class performance of the optimal texture feature configuration with the optimal configurations of the other features (corresponding to Table I). Ignoring the baseline simple statistics, the texture features perform the best on three and the worst on one of the classes. They perform well on the beach and golf course classes but poorly on the chaparral class.

Fig. 6 shows the “confusion matrices” for the different features. The rows indicate the query class and the columns the target classes. For each query image, we record the fraction of images in the top 100 retrievals that are in each of the 21 target classes. These values are then averaged over all 100 query images and displayed in the confusion matrices using a grayscale colorbar. For example, a value of 0.72 in row X and column Y indicates that on average, 72% of the top 100 retrievals had class Y when class X is the query image. The confusion matrix in Fig. 6(b) indicates the forest, river,

TABLE III
PERFORMANCE OF COLOR HISTOGRAM FEATURES

	HLS		Lab		RGB	
Dissimilarity Measure	**ANMRR**	**Time (sec.)**	**ANMRR**	**Time (sec.)**	**ANMRR**	**Time (sec.)**
Bhattacharyya	0.7490	37.15	0.7433	37.29	0.7446	34.77
Chi-Square	0.7447	186.7	*0.7405*	134.5	*0.7402*	149.2
Correlation	0.7769	15.80	0.7702	15.68	0.7711	15.64
Cosine	0.7765	15.49	0.7700	15.49	0.7711	15.50
EMD	0.7371	1.650E+05	0.7414	5.820E+03	0.7453	1.417E+04
EMD Cost Matrix	*0.7351*	1.654E+05	0.7414	5.813E+03	0.7453	1.445E+04
Inner Product	0.8005	6.920	0.8335	6.880	0.8016	6.850
Intersection	0.7468	13.99	0.7455	14.06	0.7438	14.01
L1	0.7468	5.56	0.7455	5.56	0.7438	5.57
L2	0.7894	6.14	0.7648	6.13	0.7789	6.13

and three residential classes appear very similar to the chaparral class with respect to the texture features.

B. Color Histogram

Table III lists the performance of the color histogram features computed in the three different color spaces and compared using different dissimilarity measures. While the performance varies with the particular color space and measure, it is overall much worse than the texture features (compare with Table II) and only marginally better than the baseline simple statistics. This is not unexpected as many of the classes are spectrally similar and differ mostly spatially. The best results use the EMD distance with a cost matrix applied in the HLS color space although the increased computational complexity is evident in the timing.

Interestingly, there is no best color space. The HLS color space is optimal for the two variants of the EMD and the inner product dissimilarity measures; the Lab color space is optimal for the Bhattacharyya, correlation, cosine, and L2 dissimilarity measures; and the RGB color space is optimal for the chi-square, intersection, and L1 dissimilarity measures.

Likewise, there also is no best dissimilarity measure. The correlation, cosine, and inner product measures generally perform poorly. The chi-square measure is optimal for the Lab and RGB color spaces, and second only to the EMD measures for the HLS color space, and thus could be considered the best measure overall. The L1 measure is computationally efficient while nearly optimal. (While the intersection measure is equal to the L1 measure for L1 normalized histograms, the L1 measure is computationally more efficient since it avoids computing the minimum between feature components.) The chi-square and L1 measures thus provide a performance-efficiency tradeoff for similarity retrieval using color histogram features.

Fig. 5 indicates the optimal color histogram configuration does not perform the best on any and performs the worst on 14 of the classes (again, ignoring the baseline simple statistics). Color histogram features perform poorly on the agricultural, freeway, and runway classes. The confusion matrix in Fig. 6(c) indicates that, with respect to color histogram features, the agricultural class appears similar to the golf course class, the freeway class appears similar to the intersection and overpass classes, and the runway class appears similar to the airport and freeway classes. This makes sense based on the sample images in Fig. 2.

C. Local Features

This section first presents some of the more general observations on the performance of the local invariant features, such as whether the codebooks should be constructed using k-means clustering based on the Euclidean or Mahalanobis distance. It then focuses on details such as saliency- versus grid-based feature extraction, the effect of the codebook size, and the choice of dissimilarity measure. Finally, the local features are compared to the other features considered in this study.

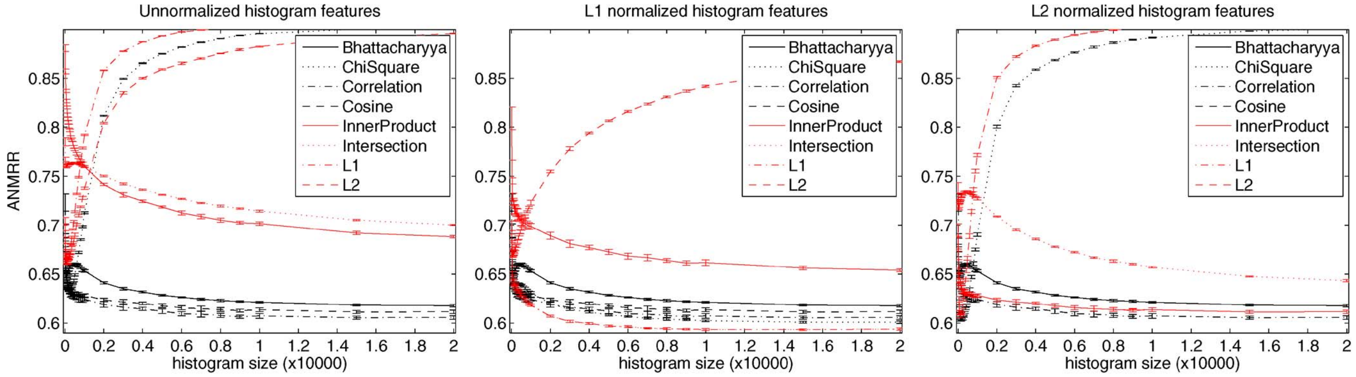
An exhaustive set of experiments are performed using all possible combinations of codebook construction, feature normalization, and dissimilarity measures. Codebooks were constructed through k-means clustering using either 100 000 or one million randomly sampled SIFT descriptors using either the Euclidean or Mahalanobis distance, and for codebook sizes ranging from 10 to 20 000. Ten codebooks were created in each case by randomly sampling different sets of SIFT descriptors. Local feature histograms with dimension equal to the the size of the codebooks were computed for each image based on the unnormalized counts of the quantized features. Histograms of L1 and L2 normalized counts were also computed. Finally, dissimilarity comparison was computed using the Bhattacharyya, chi-square, correlation, cosine, inner product, intersection, L1, L2, and EMD cost matrix measures.

Clustering Using the Euclidean Versus Mahalanobis Distance: The experiments indicate that codebooks constructed through k-means clustering using the Euclidean distance perform better than those constructed using the Mahalanobis distance. The Euclidean distance codebooks result in a 1–2% increase in performance on average independent of all other settings: sample size, codebook size, feature normalization, and dissimilarity measure. Thus, it appears that the correlations between dimensions in the 128-dimensional SIFT feature space as well as the difference in scales along the dimensions is important when using k-means clustering to construct the codebooks.

Clustering One Hundred Thousand Versus One Million Points: Codebooks constructed by applying k-means clustering to a million randomly sampled SIFT descriptors similarly perform better than those constructed using only 100 thousand descriptors. The increase is again around 1–2% on average independent of other settings and can be as high as 5%. Thus, the additional computation of applying k-means to larger sample sets of points is worthwhile particularly since this is a preprocessing step which does not impact the cost of retrieval.

TABLE IV
ANMRR VALUES FOR SALIENCY- AND GRID-BASED LOCAL FEATURE EXTRACTION. VALUES REPORTED CORRESPOND TO THE OPTIMAL CODEBOOK SIZE AND NORMALIZATION SCHEME (NOT LISTED) FOR EACH DISSIMILARITY MEASURE

	Feature extraction			
Dissimilarity Measure	**Saliency**	**4-pixel grid**	**8-pixel grid**	**16-pixel grid**
Bhattacharyya	0.6178 ± 1.254E-3	0.6326 ± 2.218E-3	0.6461 ± 1.919E-3	0.6688 ± 2.808E-3
Chi-Square	0.6009 ± 1.209E-3	0.6201 ± 2.240E-3	0.6262 ± 3.138E-3	0.6604 ± 6.505E-3
Correlation	0.6055 ± 1.635E-3	0.6430 ± 2.687E-3	0.6679 ± 3.806E-3	0.6964 ± 5.931E-3
Cosine	0.6113 ± 1.665E-3	0.6416 ± 2.045E-3	0.6605 ± 2.103E-3	0.6888 ± 1.783E-3
Inner Product	0.6113 ± 1.665E-3	0.6423 ± 1.814E-3	0.6605 ± 2.103E-3	0.6888 ± 1.783E-3
Intersection	0.5933 ± 9.076E-4	0.6201 ± 2.298E-3	0.6310 ± 1.998E-3	0.6647 ± 2.157E-3
L1	0.5933 ± 9.076E-4	0.6201 ± 2.278E-3	0.6310 ± 1.986E-3	0.6644 ± 2.171E-3
L2	0.6113 ± 1.665E-3	0.6423 ± 1.814E-3	0.6605 ± 2.103E-3	0.6888 ± 1.783E-3



(a) (b) (c)

Fig. 7. Effect of codebook size on retrieval performance for different dissimilarity measures. Results are shown for (a) unnormalized histogram features, (b) L1 normalized histogram features, and (c) L2 normalized histogram features. These results are for saliency-based feature extraction.

Saliency-Based Versus Dense Extraction: The remainder of the results in this section assume codebooks constructed by applying k-means clustering using the Euclidean distance to one million randomly sampled SIFT descriptors.

The experiments indicate that local invariant features extracted from salient image locations outperform on average those extracted on a grid. Table IV compares the performance of saliency- to grid-based feature extraction for different grid spacings. Saliency-based extraction is shown to be optimal for all dissimilarity measures. This reconfirms the benefit of extracting features at locations based on the image content that was the original motivation behind the SIFT and similar detectors. The SIFT detector is designed to identify the same object components regardless of where they appear in the image and thus is not affected by possible misalignment problems that result from using a fixed grid. Further, saliency-based extraction only considers image locations where there is meaningful image information. We note that other researchers have found the opposite, that grid-based extraction is optimal. However, this was for image classification and not retrieval and was not for geographic images. Per-class comparisons between the optimal saliency- and grid-based configuration are shown in Fig. 5. The optimal grid-based configuration was found to be using the L1 dissimilarity measure to compare unnormalized histograms created using 20 000 codewords extracted from a 4-pixel grid. The best ANMRR for this configuration was 0.6179.

Codebook Size and Dissimilarity Measure: Fig. 7 summarizes the performance of the different dissimilarity measures by plotting ANMRR values for codebook sizes ranging from 10 to 20 000. These results correspond to saliency-based feature extraction—the plots have similar shapes for grid-based feature extraction but are shifted up (worse ANMRR). Three plots are shown. Fig. 7(a) shows the results for unnormalized histogram features, Fig. 7(b) for L1 normalized histogram features, and Fig. 7(c) for L2 normalized histogram features. Error bars indicate the standard deviation of the ANMRR values over the ten different codebooks (again, corresponding to clustering different random sets of SIFT descriptors). The results corresponding to the EMD cost matrix measure were significantly worse than any other measure and are thus not included in the detailed analysis.

The best results correspond to using the L1 measure to compare L1 normalized features for larger codebook sizes. Much more can be observed from these plots however. Interestingly, the effect of codebook size on performance generally falls into two categories. Either the performance improves with increasing size, in most cases in a monotonic fashion until gently peaking for large codebooks of size around 15 000; or the performance improves sharply for small codebook sizes but then decreases steadily for larger sizes. The behavior of a specific dissimilarity measure depends on the feature normalization. These findings are significant because they show that *the range of optimal codebook sizes depends greatly on the choice of normalization and measure*. Larger codebooks tend to result in increased performance. However, there are some important cases where a narrow range of small codebook sizes is nearly optimal which is important since retrieval and storage costs are often a concern.

TABLE V
Results for Different Dissimilarity Measures for Histograms of Quantized Local Features. These Results Are for Saliency-Based Feature Extraction

Dissimilarity Measure	**ANMRR (best)**	**# Codewords**	**Time (sec.)**	**ANMRR (overall)**	**Normalization**
Bhattacharyya	0.6161	20000	1518	0.6178 ± 1.254E-3	Unnormalized
Chi-Square	0.5989	20000	835.1	0.6009 ± 1.209E-3	L1
Correlation	0.6018	15000	439.5	0.6055 ± 1.635E-3	Unnormalized
Cosine	0.6070	15000	432.3	0.6113 ± 1.665E-3	Unnormalized
Inner Product	0.6070	15000	201.3	0.6113 ± 1.665E-3	L2
Intersection	*0.5914*	15000	377.8	0.5933 ± 9.076E-4	L1
L1	*0.5914*	15000	193.3	0.5933 ± 9.076E-4	L1
L2	0.6070	15000	197.4	0.6113 ± 1.665E-3	L2
Chi-Square	0.6014	150	13.43	0.6034 ± 1.510E-3	L2
L1	0.6045	150	1.770	0.6068 ± 1.288E-3	L2

The best performance for unnormalized features results from the correlation measure applied to a codebook of size 15 000 [see Fig. 7(a)]. This performance is not significant since it is still worse than the best results corresponding to normalized features, and correlation is one of the more computationally expensive measures. The performance of the L1, L2, and chi-square measures does peak for small codebook sizes but not sufficiently to make the application of these measures to unnormalized features a competitive configuration.

The best performance for L1 normalized features results from the L1 or, equivalently, intersection measure applied to a codebook of size 15 000 [see Fig. 7(b)]. The performance of the L2 distance does peak for small codebook sizes but again not sufficiently.

The best performance for L2 normalized features results from the chi-square measure applied to a codebook of size only 150 [see Fig. 7(c)]. The chi-square performance peaks for a narrow range of small codebook sizes. The next best performance is the correlation measure applied to a codebook of size 15 000. Also, noteworthy is the L1 measure which also peaks for small codebook sizes and whose optimal performance for a codebook of size 150 is only slightly worse than the optimal chi-square and correlation results but is significantly more efficient.

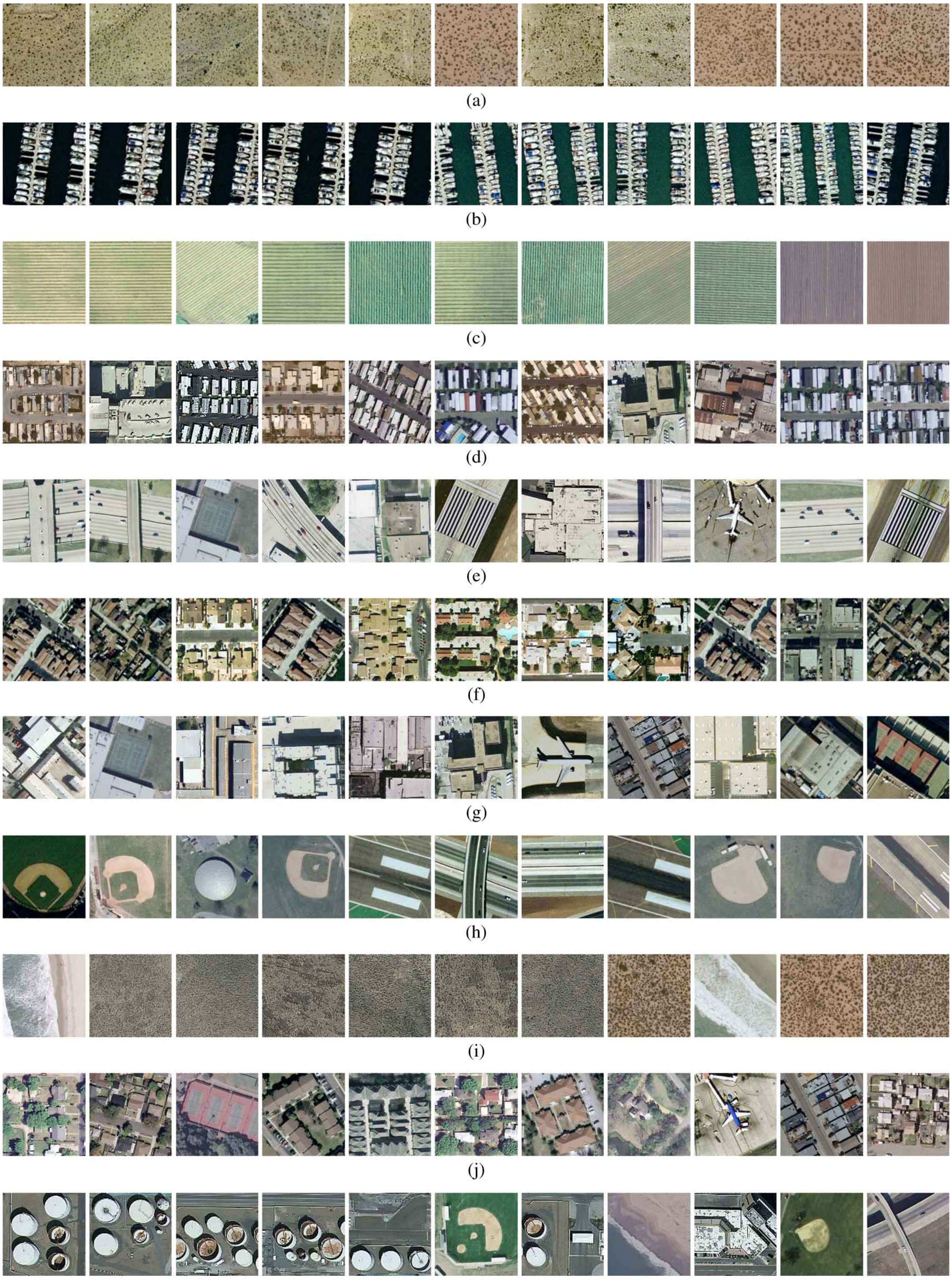
Table V shows the optimal performance for each of the dissimilarity measures. Since each measure/codebook-size/feature-normalization combination is evaluated using ten codebooks corresponding to different random sets of SIFT descriptors, this table reports both the results from the best codebook in the column labeled "ANMRR (best)" as well as the average and standard deviation over the ten codebooks in the column labeled "ANMRR (overall)." The ranking of the dissimilarity measures is the same for both. Also, shown in the last two rows is the results for the chi-square and L1 measures for small codebooks of size 150 of L2 normalized histogram features. These configurations perform slightly worse than the optimal configuration resulting in performance reductions of 1.7% and 2.2%, respectively. However, they are significantly more efficient, requiring histogram features that are two orders of magnitude smaller and retrieval times that are one and two orders of magnitude faster, respectively. They are still more effective and more efficient than the optimal configurations of the texture and color histogram features shown in Table I, and thus represent an excellent efficiency-performance trade off.

Fig. 5 indicates the optimal local feature configuration for sparse-based extraction performs the best on eight and the worst on five of the classes. This configuration performs particularly well on the chaparral, harbor, and parking classes when compared to the other methods. The optimal local feature configuration for grid-based extraction performs the best on ten and the worst on one of the classes (but still performs worse than sparse-based extraction when averaged across all classes). This configuration performs particularly well on the forest class when compared to the other methods.

Fig. 8 presents select retrieval results corresponding to the optimal local feature configuration for sparse-based extraction. The results are shown for queries from 11 classes ordered by decreasing performance based on average ANMRR.

VII. Discussion

It is worthwhile discussing the performance of the local features in the context of the desirable properties described in Section III-A. The experimental results indicate the saliency-based local features perform well on the chaparral, dense residential, harbor, mobile home park, and parking classes. These classes are characterized by repeated occurrences of a specific object—i.e., parking lots consist of cars parked in varying spatial arrangements—which leads to hypothesize the following. The *local* property enables the features to detect the individual object instances as opposed to characterizing the gestalt or overall essence of an image. The *invariance* property enables the objects to be detected regardless of where or in what orientation they appear in an image. It also enables them to be detected independent of photometric variations such as illumination intensity and density as well as differences in color. The features are shown to be *robust* to other image variations such as the amount of the noise in the images and the quality of the different camera systems in terms of visual acuity (see, for example, the variation in the images of the parking lot class in Fig. 2). The *density* of the features allows them to be robust against occlusion caused by shadows, trees, etc. It also allows them to be robust against missed detections which is important for detecting the large number of object occurrences. The local features are *efficient* which is import for real-time application or use in large image data sets. The appropriate pairing of small codebooks and dissimilarity measures resulted in retrieval performance times that were an order of magnitude or faster than competitive features. Moreover, while we did not



(a)

(b)

(c)

(d)

(e)

(f)

(g)

(h)

(i)

(j)

(k)

Fig. 8. Sample retrievals for different classes corresponding to the optimal local feature configuration for sparse-based extraction. The classes are ordered in increasing average ANMRR. The leftmost image is the query image in each case, and the remaining images are the top ten retrievals in decreasing order of similarity. The captions indicate the average ANMRR for the classes as well as the rank and class of the incorrect retrievals for the specific query. (a) Chaparral: ANMRR $= 0.0316$; (b) harbor: average ANMRR $= 0.2757$; (c) agricultural: ANMRR $= 0.3537$; (d) mobile home park: ANMRR $= 0.5202$ (1: buildings; 3: dense residential; 7: buildings; 8: buildings); (e) overpass: ANMRR $= 0.5786$ (2: tennis courts; 3: freeway; 4: buildings; 5: runway; 6: buildings; 8: airplane; 9: freeway; 10: runway); (f) dense residential: ANMRR $= 0.6$ (6: medium density residential; 7: medium density residential; 9: intersection); (g) buildings: ANMRR $= 0.6771$ (1: tennis courts; 6: airplane; 7: dense residential; 10: tennis courts); (h) baseball diamond: ANMRR $= 0.7507$ (2: storage tanks; 4: runway; 5 overpass; 6: freeway; 7: runway; 10 runway); (i) beach: ANMRR $= 0.7828$ (1: chaparral; 2: chaparral; 3: chaparral; 4: chaparral; 5: chaparral; 6: chaparral; 7: chaparral; 9: chaparral; 10: chaparral); (j) medium density residential: ANMRR $= 0.7998$ (2: tennis courts; 3: dense residential; 4: dense residential; 6: dense residential; 7: sparse residential; 8: airplane; 9: dense residential); (k) storage tanks: ANMRR $= 0.8451$ (5: baseball diamond; 7: beach; 8: buildings; 9: baseball diamond; 10: overpass).

address the computational costs of feature extraction in this paper, we demonstrated in earlier work that SIFT features are a magnitude faster to extract than Gabor texture features [59].

The saliency-based local features perform poorly on the baseball diamond, beach, golf, and runway classes. Saliency-based feature extraction results in relatively few features in these classes as indicated in Fig. 3 as they tend to contain large uniform regions. The sparseness of the resulting BOVW histograms reduces their discrimination ability.

The texture features perform well on the beach, baseball diamond, golf course, intersection, river, and the sparse and medium density residential classes. The grid-based local features also perform well on these classes. This correlation is likely due to the fact that extracting SIFT descriptors—which are after all summaries of local edge information—on a regular grid is similar to applying Gabor filters which can also be considered local edge detectors.

Finally, it is interesting that the saliency-based local features perform better than the texture features on the classes which upon first inspection appear more "texture" like. For example, the rows of cars in the parking lots or the boats in the harbor result in the kinds of regular patterns suitable for representation by frequency-based texture features. However, upon closer examination, these patterns do vary at the microscopic level in that the elements are often missing, such as empty boat slips, or are arranged at different angles with respect to each other such as the parked cars. The patterns also vary at the macroscopic level in that the rows do not necessarily have the same spacing from one image to another. Saliency-based local features are more invariant to these micro- and macroscopic variations since they instead detect the individual objects or parts thereof. These irregularities can also explain why the grid-based local features perform poorly on these classes.

VIII. Conclusion

We presented an investigation into local invariant features for overhead image retrieval, the first such study of its kind. We demonstrated that local invariant features are more effective than standard features such as color and texture for image retrieval of LULC classes in high-resolution aerial imagery. We also quantitatively analyzed the effects of a number of design parameters on a BOVW representation including saliency- versus grid-based feature extraction, the size of the visual codebook, the clustering algorithm used to create the codebook, and the dissimilarity measure used to compare the BOVW representations. We feel such a study is timely given the increased interest by the remote sensing community in using local features for image analysis. While the focus is on image retrieval, we expect the insights on the effects of the design parameters to be informative for other applications such as detection and classification.

We created a first-of-its-kind 21 LULC evaluation data set using imagery that is already in the public domain. We will make this data set available to other researchers with the expectation that it will help advance overhead image analysis in the same way that similar data sets have done for other areas of computer vision.

Acknowledgment

The authors would like to thank the anonymous reviewers for their comments and corrections.

References

[1] H. Goncalves, L. Corte-Real, and J. Goncalves, "Automatic image registration through image segmentation and SIFT," *IEEE Trans. Geosci. Remote Sens.*, vol. 49, no. 7, pp. 2589–2600, Jul. 2011.

[2] A. Sedaghat, M. Mokhtarzade, and H. Ebadi, "Uniform robust scale-invariant feature matching for optical remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 49, no. 11, pp. 4516–4527, Nov. 2011.

[3] X. Jianbin, H. Wen, and W. Yirong, "An efficient rotation-invariance remote image matching algorithm based on feature points matching," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2005, vol. 1, pp. 647–649.

[4] J. Dai, W. Song, L. Pei, and J. Zhang, "Remote sensing image matching via Harris detector and SIFT discriptor," in *Proc. Int. Congr. Image Signal Process.*, 2010, vol. 5, pp. 2221–2224.

[5] A. Mukherjee, M. Velez-Reyes, and B. Roysam, "Interest points for hyperspectral image data," *IEEE Trans. Geosci. Remote Sens.*, vol. 47, no. 3, pp. 748–760, Mar. 2009.

[6] L. Dorado-Munoz, M. Velez-Reyes, A. Mukherjee, and B. Roysam, "A vector SIFT operator for interest point detection in hyperspectral imagery," in *Proc. Workshop Hyperspectr. Image Signal Process.—Evolution Remote Sensing*, 2010, pp. 1–4.

[7] Z. Xiong and Y. Zhang, "A novel interest-point-matching algorithm for high-resolution satellite images," *IEEE Trans. Geosci. Remote Sens.*, vol. 47, no. 12, pp. 4189–4200, Dec. 2009.

[8] C. Huo, Z. Zhou, Q. Liu, J. Cheng, H. Lu, and K. Chen, "Urban change detection based on local features and multiscale fusion," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2008, vol. 3, pp. 1236–1239.

[9] F. Tang and V. Prinet, "Computing invariants for structural change detection in urban areas," in *Proc. Urban Remote Sens. Joint Event*, 2007, pp. 1–6.

[10] B. Sirmacek and C. Unsalan, "Urban-area and building detection using SIFT keypoints and graph theory," *IEEE Trans. Geosci. Remote Sens.*, vol. 47, no. 4, pp. 1156–1167, Apr. 2009.

[11] B. Sirmacek and C. Unsalan, "Urban area detection using local feature points and spatial voting," *IEEE Geosci. Remote Sens. Lett.*, vol. 7, no. 1, pp. 146–150, Jan. 2010.

[12] B. Sirmacek and C. Unsalan, "A probabilistic framework to detect buildings in aerial and satellite images," *IEEE Trans. Geosci. Remote Sens.*, vol. 49, no. 1, pp. 211–221, Jan. 2011.

[13] S. Xu, T. Fang, D. Li, and S. Wang, "Object classification of aerial images with bag-of-visual words," *IEEE Geosci. Remote Sens. Lett.*, vol. 7, no. 2, pp. 366–370, Apr. 2010.

[14] L. Chen, W. Yang, K. Xu, and T. Xu, "Evaluation of local features for scene classification using VHR satellite images," in *Proc. Urban Remote Sens. Joint Event*, 2011, pp. 385–388.

[15] A. Skurikhin, "Visual attention based detection of signs of anthropogenic activities in satellite imagery," in *Proc. IEEE Appl. Imag. Pattern Recog. Workshop*, 2010, pp. 1–8.

[16] S. Gleason, R. Ferrell, A. Cheriyadat, R. Vatsavai, and S. De, "Semantic information extraction from multispectral geospatial imagery via a flexible framework," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2010, pp. 166–169.

[17] R. R. Vatsavai, A. Cheriyadat, and S. Gleason, "Unsupervised semantic labeling framework for identification of complex facilities in high-resolution remote sensing images," in *Proc. Int. Conf. Data Mining Workshops*, 2010, pp. 273–280.

[18] B. Ozdemir and S. Aksoy, "Image classification using subgraph histogram representation," in *Proc. Int. Conf. Pattern Recog.*, 2010, pp. 1112–1115.

[19] J. Bordes and V. Prinet, "Mixture distributions for weakly supervised classification in remote sensing images," in *Proc. Brit. Mach. Vis. Conf.*, 2008.

[20] L. Fei-Fei, R. Fergus, and P. Perona, "Learning generative visual models from few training examples: An incremental Bayesian approach tested on 101 object categories," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recog*, 2004, p. 178.

[21] G. Griffin, A. Holub, and P. Perona, "Caltech-256 object category dataset," California Inst. Technol., Pasadena, CA, Tech. Rep. 7694, 2007.

[22] M. Everingham, A. Zisserman, C. Williams, L. Van Gool, M. Allan, C. Bishop, O. Chapelle, N. Dalal, T. Deselaers, G. Dorkó, S. Duffner, J. Eichhorn, J. Farquhar, M. Fritz, C. Garcia, T. Griffiths, F. Jurie, D. Keysers, M. Koskela, J. Laaksonen, D. Larlus, B. Leibe, H. Meng,

H. Ney, B. Schiele, C. Schmid, E. Seemann, J. Shawe-Taylor, A. Storkey, S. Szedmak, B. Triggs, I. Ulusoy, V. Viitaniemi, and J. Zhang, "The 2005 PASCAL visual object classes challenge," in *Proc. Pascal Challenges Workshop*, vol. 3944, *Lecture Notes in Computer Science*, 2006, pp. 117–176.

[23] J. Ashley, M. Flickner, J. Hafner, D. Lee, W. Niblack, and D. Petkovic, "The query by image content (QBIC) system," in *Proc. ACM SIGMOD Int. Conf. Manag. Data*, 1995, p. 475.

[24] Q. Bao and P. Guo, "Comparative studies on similarity measures for remote sensing image retrieval," in *Proc. IEEE Int. Conf. Syst., Man Cybern.*, 2004, pp. 1112–1116.

[25] T. Bretschneider, R. Cavet, and O. Kao, "Retrieval of remotely sensed imagery using spectral information content," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2002, pp. 2253–2255.

[26] T. Bretschneider and O. Kao, "A retrieval system for remotely sensed imagery," in *Proc. Int. Conf. Imag. Sci., Syst., Technol.*, 2002, vol. 2, pp. 439–445.

[27] G. Scott, M. Klaric, C. Davis, and C.-R. Shyu, "Entropy-balanced bitmap tree for shape-based object retrieval from large-scale satellite imagery databases," *IEEE Trans. Geosci. Remote Sens.*, vol. 49, no. 5, pp. 1603–1616, May 2011.

[28] A. Ma and I. K. Sethi, "Local shape association based retrieval of infrared satellite images," in *Proc. IEEE Int. Symp. Multimedia*, 2005, pp. 551–557.

[29] M. Ferecatu and N. Boujemaa, "Interactive remote-sensing image retrieval using active relevance feedback," *IEEE Trans. Geosci. Remote Sens.*, vol. 45, no. 4, pp. 818–826, Apr. 2007.

[30] B. Raghunathan and S. Acton, "Content based retrieval for remotely sensed imagery," in *Proc. IEEE Southwest Symp. Image Anal. Interpretation*, 2000, pp. 161–165.

[31] K. W. Tobin, B. L. Bhaduri, E. A. Bright, A. Cheriyadat, T. P. Karnowski, P. J. Palathingal, T. E. Potok, and J. R. Price, "Large-scale geospatial indexing for image-based retrieval and analysis," in *Proc. Int. Symp. Vis. Comput.*, 2005, pp. 543–552.

[32] K. W. Tobin, B. L. Bhaduri, E. A. Bright, A. Cheriyadat, T. P. Karnowski, P. Palathingal, T. E. Potok, and J. R. Price, "Automated feature generation in large-scale geospatial libraries for content-based indexing," *Photogramm. Eng. Remote Sens.*, vol. 72, no. 5, pp. 531–540, May 2006.

[33] Y. Li and T. Bretschneider, "Semantics-based satellite image retrieval using low-level features," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2004, vol. 7, pp. 4406–4409.

[34] Y. Hongyu, L. Bicheng, and C. Wen, "Remote sensing imagery retrieval based-on Gabor texture feature classification," in *Proc. Int. Conf. Signal Process.*, 2004, pp. 733–736.

[35] B. S. Manjunath and W. Y. Ma, "Texture features for browsing and retrieval of image data," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 18, no. 8, pp. 837–842, Aug. 1996.

[36] 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," *J. Appl. Opt.*, vol. 43, no. 2, pp. 210–217, Jan. 2004.

[37] A. Samal, S. Bhatia, P. Vadlamani, and D. Marx, "Searching satellite imagery with integrated measures," *Pattern Recognit.*, vol. 42, no. 11, pp. 2502–2513, Nov. 2009.

[38] S. Newsam and C. Kamath, "Retrieval using texture features in high resolution multi-spectral satellite imagery," in *Proc. SPIE Defense Security Symp., Data Mining Knowl. Discov.: Theory, Tools, Technol. VI*, 2004, pp. 21–32.

[39] C. Harris and M. Stephens, "A combined corner and edge detector," in *Proc. 4th Alvey Vis. Conf.*, 1988, pp. 147–151.

[40] K. Mikolajczyk and C. Schmid, "Scale & affine invariant interest point detectors," *Int. J. Comput. Vis.*, vol. 60, no. 1, pp. 63–86, Oct. 2004.

[41] T. Lindeberg, "Feature detection with automatic scale selection," *Int. J. Comput. Vis.*, vol. 30, no. 2, pp. 79–116, Nov. 1998.

[42] K. Mikolajczyk and C. Schmid, "A performance evaluation of local descriptors," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 10, pp. 1615–1630, Oct. 2005.

[43] D. Nister and H. Stewenius, "Scalable recognition with a vocabulary tree," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recog.*, 2006, vol. 2, pp. 2161–2168.

[44] Snaptell—Visual Product Search. [Online]. Available: http://snaptell.com/

[45] D. G. Lowe, "Object recognition from local scale-invariant features," in *Proc. IEEE Int. Conf. Comput. Vis.*, 1999, vol. 2, pp. 1150–1157.

[46] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, Nov. 2004.

[47] A. Z. T. Kadir and M. Brady, "An affine invariant salient region detector," in *Proc. Eur. Conf. Comput. Vis.*, 2004, pp. 404–416.

[48] S. Belongie, J. Malik, and J. Puzicha, "Shape matching and object recognition using shape contexts," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 4, pp. 509–522, Apr. 2002.

[49] W. T. Freeman and E. H. Adelson, "The design and use of steerable filters," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 13, no. 9, pp. 891–906, Sep. 1991.

[50] Y. Ke and R. Sukthankar, "PCA-SIFT: A more distinctive representation for local image descriptors," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recog.*, 2004, vol. 2, pp. 506–513.

[51] S. Lazebnik, C. Schmid, and J. Ponce, "Sparse texture representation using affine-invariant neighborhoods," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recog.*, 2003, pp. 319–324.

[52] L. J. V. Gool, T. Moons, and D. Ungureanu, "Affine/photometric invariants for planar intensity patterns," in *Proc. Eur. Conf. Comput. Vis.*, 1996, pp. 642–651.

[53] J. Matas, O. Chum, M. Urban, and T. Pajdla, "Robust wide baseline stereo from maximally stable extremal regions," in *Proc. Brit. Mach. Vis. Conf.*, 2002, vol. 1, pp. 384–393.

[54] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, "Speeded-up robust features (SURF)," *Comput. Vis. Image Understand.*, vol. 110, no. 3, pp. 346–359, Jun. 2008.

[55] J. Sivic and A. Zisserman, "Video Google: A text retrieval approach to object matching in videos," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2003, vol. 2, pp. 1470–1477.

[56] B. S. Manjunath, P. Salembier, and T. Sikora, Eds., *Introduction to MPEG7: Multimedia Content Description Interface*. New York: Wiley, 2002.

[57] Y. Rubner, C. Tomasi, and L. Guibas, "A metric for distributions with applications to image databases," in *Proc. IEEE Int. Conf. Comput. Vis.*, 1998, pp. 59–66.

[58] P. Wu, B. Manjunath, S. Newsam, and H. Shin, "A texture descriptor for browsing and similarity retrieval," *J. Signal Process., Image Commun.*, vol. 16, no. 1/2, pp. 33–43, Sep. 2000.

[59] Y. Yang and S. Newsam, "Comparing SIFT descriptors and Gabor texture features for classification of remote sensed imagery," in *Proc. IEEE Int. Conf. Image Process.*, 2008, pp. 1852–1855.

Yi Yang (S'08) received the B.E. degree in control engineering from Tsinghua University, Beijing, China, in 2003. He is currently working toward the Ph.D. degree in the electrical engineering and computer science program at the University of California, Merced.

His main research is focused on remotely sensed image analysis, including geographic image retrieval, land-use/land-cover classification, and integration of image and non-image geospatial data.

Shawn Newsam (M'00) received the B.S. degree in electrical engineering and computer science from the University of California, Berkeley, in 1990, the M.S. degree in electrical and computer engineering from the University of California, Davis, in 1996, and the Ph.D. degree in electrical and computer engineering from the University of California, Santa Barbara, in 2004.

Currently, he is an Assistant Professor and founding faculty of electrical engineering and computer science at the University of California, Merced. Prior to joining U.C. Merced, he was a Postdoctoral Scholar in the Center for Applied Scientific Computation at the Lawrence Livermore National Laboratory. His research interests are in image processing, computer vision, and pattern recognition particularly as applied to multidisciplinary problems.

Dr. Newsam is the Recipient of a U.S. Department of Energy Early Career Scientist and Engineer Award, a U.S. National Science Foundation Faculty Early Career Development (CAREER) Award, and a U.S. Office of Science and Technology Policy Presidential Early Career Award for Scientists and Engineers.