Spatial Feature Evaluation for Aerial Scene Analysis

Thomas Swearingen,
Student Member, *IEEE*
University of Tennessee, Knoxville
tsweari1@utk.edu

Anil Cheriyadat,
Member, *IEEE*
Oak Ridge National Laboratory
cheriyadatam@ornl.gov

***Abstract*—High-resolution aerial images are becoming more readily available, which drives the demand for robust, intelligent and efficient systems to process increasingly large amounts of image data. However, automated image interpretation still remains a challenging problem. Robust techniques to extract and represent features to uniquely characterize various aerial scene categories is key for automated image analysis. In this paper we examined the role of spatial features to uniquely characterize various aerial scene categories. We studied low-level features such as colors, edge orientations, and textures, and examined their local spatial arrangements. We computed correlograms representing the spatial correlation of features at various distances, then measured the distance between correlograms to identify similar scenes. We evaluated the proposed technique on several aerial image databases containing challenging aerial scene categories. We report detailed evaluation of various low-level features by quantitatively measuring accuracy and parameter sensitivity. To demonstrate the feature performance, we present a simple query-based aerial scene retrieval system.**

***Index Terms*—Aerial scene, low-level features, spatial correlation, retrieval, classification**

I. Introduction

Robust techniques that can generate numerical representations to uniquely characterize various aerial scene categories while remaining invariant to changes in appearance attributes and geometrical transformations are in great demand today. Such techniques play a vital role in various image analysis systems developed for automated scene recognition, content-based image indexing, and change detection. In the past, researchers have explored several interesting approaches based on pixel and object (homogeneous pixel groups) features for aerial image classification [1], [2], [3], [4], [5]. Often, these techniques are limited in the way they exploit the rich scene attributes offered by the high-resolution aerial scenes. Another alternative is to segment the scene into components that correspond to the underlying physical objects (such as roads, parking lots, or buildings) in the scene, then generate representations based on the object frequencies and spatial arrangements. However, accurate image segmentation still remains a challenging problems making these approaches less viable and computationally demanding.

In the case of high-resolution aerial scenes, the difficulties in generating unique representations for scene categories are further compounded by the high within-class scene variations and the low between-class variations. For example, in the case of *mobile-home park* scenes, the appearance variations of the mobile homes and the differences in their spatial arrangements itself can give rise to the wide within-class variations as shown by the example images provided in the top row of figure 1. Additionally, other scene categories such as the *warehouse* or *trailer parks* might display visual attributes that are similar to the *mobile-home park* scenes. Encoding the local structural attributes and their spatial patterns in an effective way is key for generating robust scene representations. In this paper we propose a representation technique that captures the local structural and spatial scene attributes in an efficient manner to characterize various aerial scene categories. Our approach can be considered as a trade-off between simple pixel based representation and computationally demanding segmentation based representations. The proposed method begins by breaking up the scene into local image patches, then generating low-level feature based representation for the image patches. Next, we measure the spatial arrangement of local image patches by computing correlgorams. Similarity between different aerial scenes are measured based on comparing correlograms through a unique distance metric. Previously Huang *et. al.* [6] explored similar color correlgrams for measuring image similarity.

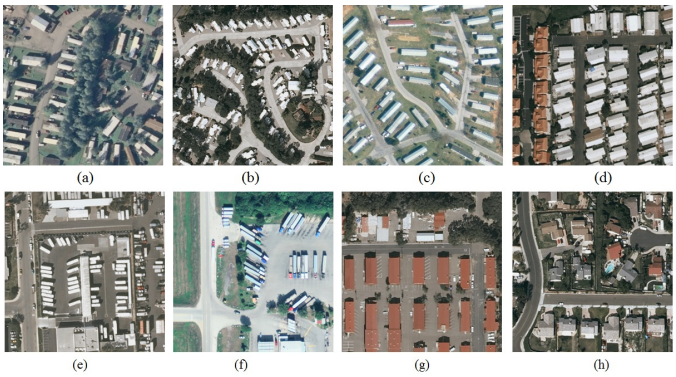

Fig. 1. Top row (a-d) shows a few example images representing mobile-home parks. The bottom row shows similar looking scenes from other categories: (e) warehouse, (f) trailer park, (g) and (h) represent residential scenes.

The rest of the paper is organized as follows. In Section II we briefly review recent and relevant work on high-resolution satellite image classification that exploits spatial features. In Section III we describe our spatial feature representation approach in detail. Details of our experiments and results are presented in sections IV and V. Section VI concludes the paper with discussions on the findings and ideas for extending

the work.

II. Related Work

We start by reviewing some of the recent works that exploit pixel-level and spatial context features for high-resolution satellite image classification. Bruzzone and Carlin [7] proposed a spatial context driven feature extraction strategy for pixel classification in high-resolution satellite images. First, image segmentation was performed at different scales. Pixel-level features are combined with geometrical features computed from the associated segment for image classification. Similarly, Shackelford and Davis [4] combined both pixel-based and object-based features to generate object-level classification of the image. In contrast to the above approaches, Unsalan and Boyer in [8] showed that intermediate representation of the scene based on local line parameters provided an effective way to represent different broad aerial scenes. The statistical measures derived from line length, contrast, and orientation distributions provided unique lower-dimensional representation for different scene categories. Similarly, in [9] Huang *et. al.* explored a similar idea based on directional lines for generating pixel features. The gray-level similarity among pixels at certain distances and orientations were calculated to determine possible direction lines. Statistics computed from the directional line length histogram associated with each pixel was used as the feature vector. However the above approaches do not measure the spatial arrangement of features and were limited in the way they exploit the rich scene attributes provided by the high-resolution aerial scenes in the image. In contrast to the above approaches our proposed method provides a holistic representation for the scene based on spatial correlograms of low-level features. Next, we describe our method in detail.

III. Proposed Technique

A. Overview

Given a image patch representing the scene, we first divide the image into overlapping square patches of size $q \times q$ with overlap parameter s. Next, we represent each patch in terms of their low-level feature descriptor. In this paper we evaluated three different low-level feature descriptors: (i) mean color vector, (ii) Scale Invariant Feature Transform (SIFT) to measure local edge patterns, and (iii) oriented filter responses to measure texture. Each feature descriptor is then quantized to reduce the feature space. We compute the spatial co-occurrence statistics of image patches to generate the spatial correlogram. Spatial correlogram represents the co-occurrence frequency of similar patches with respect to certain spatial predicates. Now, the given image is represented in terms of low-level feature correlgrams. Image similarity is measured based on a correlogram distance measure. Figure 2 shows the overview of the proposed method.

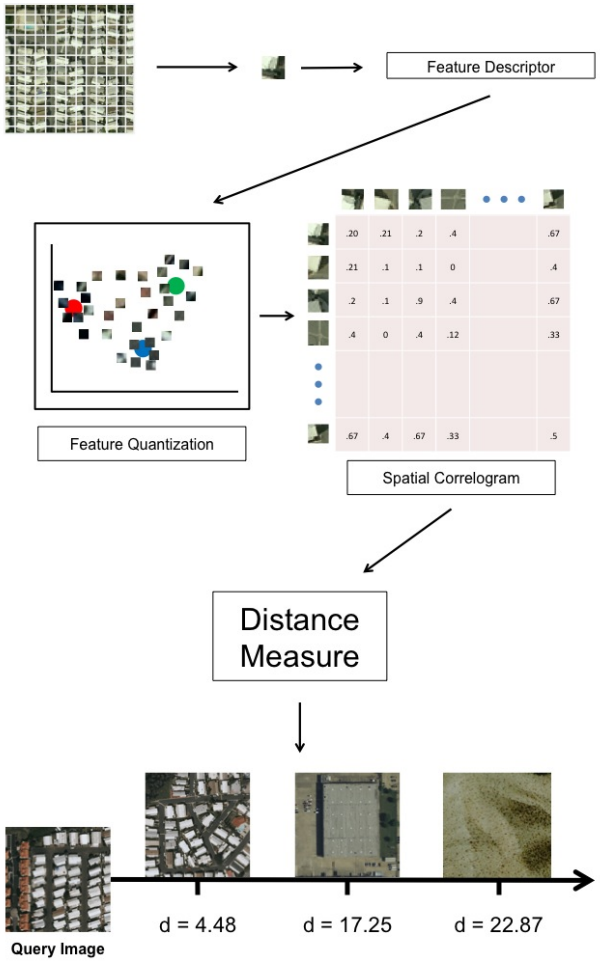

Fig. 2. Methodology overview

B. Features

1) Color: To compute color features for each patch we measure mean pixel intensity for the three different color channels. In this case each image patch is represented by a feature descriptor $f \in \mathbb{R}^3$.

2) Scale Invariant Feature Transform: To measure the edge orientation statistics we use the SIFT descriptors proposed in [10]. SIFT descriptors are invariant to small differences in the image scales, illumination conditions, rotations, and viewpoint variations. Unlike the original method in [10] which computes descriptors only at the detected interest points in the image, here we compute SIFT descriptors for each overlapping each image patch. Each image patch is represented by feature descriptor $f \in \mathbb{R}^{128}$.

3) Oriented Filter Response: Texture measure for each image patch is captured through the oriented filter responses. For oriented filter responses, we use the Leung-Malik [11] multi-scale and orientation filter banks. Our filter bank consists of first and second derivatives of Gaussian functions at 6 orientations and 3 scales, 8 Laplacian-of-Gaussian and, 4 Gaussians. Following [11], for each scale we set the Gaussian width correspondingly to $\{1, \sqrt{2}, 2, 2\sqrt{2}\}$. For each pixel patch, we compute the average filter energy at every scale and orientation to generate feature vector $f \in \mathbb{R}^{48}$.

C. Feature Quantization

Next, to reduce the number of representative feature descriptors used for measuring the spatial co-occurrence statistics we quantize the feature space by applying *k-means* clustering. We fix the number of clusters m based on experimental evaluations. The feature descriptor associated with each image patch is replaced with the associated cluster label.

D. Spatial Correlogram

Next, we measure the co-occurrence frequency of similar image patches (in this case patches with same cluster labels) at certain distances d. For every patch $p = 1, 2, \ldots, m$ we measure the co-occurrence of the patch with itself at certain pixel distances $k = 1, 2, \ldots, d$ at fixed angles of $\{0°, 90°, 180°, 270°\}$. The resulting spatial correlogram matrix M with $d \times m$ elements measures how the autocorrelation of image patches varies with distance. Each element $M(i, j)$ represents the normalized frequency of patch j co-occurring at a distance i. The normalized frequency is computed by dividing the measured co-occurrence frequency with the total possible co-occurrences for each patch (e.g. for an each image patch occurring 5 times in the image, the total possible co-occurrence is $5 \times 4 = 20$) .

E. Distance Measure

Given two images represented by correlogram matrices M_1 and M_2, we compute the image similarity based on the L_1 distance norm. As suggested in [6] we normalize the L_1 norm distance as given by equation 1. The 1 is added in the denominator of equation 1 to prevent division by zero.

$$\sum_{i \in d, j \in m} \frac{M_1(i, j) - M_2(i, j)}{1 + M_1(i, j) + M_2(i, j)} \tag{1}$$

IV. Experiments

A. Data

To test the robustness and accuracy of the spatial correlogram based aerial scene representation we applied our approach on three challenging and diverse aerial image datasets. The first dataset referred hereafter as *MobileHomePark* contains aerial images with dimensions 512×512 pixels and 0.3 meters spatial resolution. The dataset contains manually cropped images representing 3 different categories: (i) 1307 examples of mobile-home parks, (ii) 2076 examples of man-made structures other than the mobile homes, and (iii) 2340 natural scenes depicting forest and barren land areas. Some of the images in the dataset share portions of the same scene. Figure 1 shows a few example scenes from this dataset.

The second dataset referred as *UCMERCED* was compiled by [12] containing manually extracted aerial orthoimagery downloaded from the United States Geological Survey (USGS) National Map. The images have a resolution of one foot per pixel and are cropped to 256×256 pixels. The dataset contains 21 challenging scene categories with 100 samples per class. The dataset represents highly overlapping classes such as the *denseresidential*, *mediumresidential* and *sparseresidential* which mainly differs in the density of structures. Figure 3 shows examples from a few selected categories in the database.

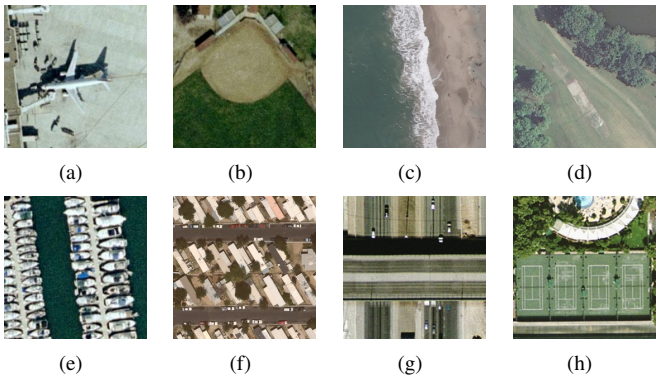

Fig. 3. Examples of select image tile categories in the *UCMERCED* dataset. The database has 21 categories: agricultural, airplane (a), baseball diamond (b), beach (c), buildings, chaparral, dense residential, forest, freeway, golf course (d), harbor (e), intersection, medium residential, mobile home park (f), overpass (g), parking lot, river, runway, sparse residential, storage tanks, and tennis courts (h).

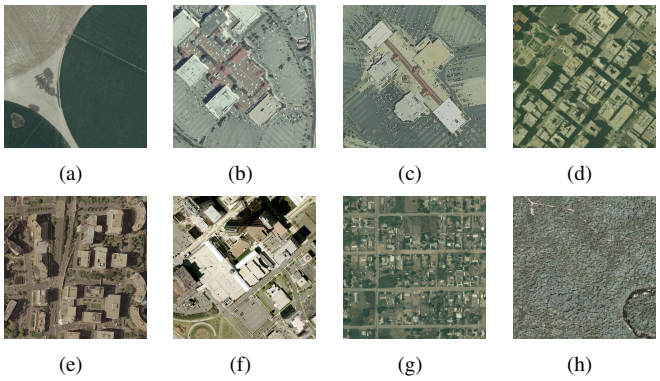

Fig. 4. Examples of image tiles in the *ORNL-I* dataset. The database has 5 categories: Agricultural (a), Large-Facility (b-c), Commercial (d-e, Suburban (f-g), and Wooded (h).

Finally we apply our approach on the *ORNL-I* dataset [13] containing approximately one meter spatial resolution satellite images representing five different geospatial neighborhood classes namely - *agricultural*, *large-facility*, *commercial*, *suburban*, and *wooded*. These images are collected from various sources including the United States Department of Agriculture's (USDA) National Agricultural Imagery Program (NAIP), Microsoft's TerraServer-USA database, and orthoimagery provided by the state of California and Utah. The collection includes 170, 153, 171, 186 and 170 images for the *agricultural*, *large-facility*, *commercial*, *suburban* and *wooded* classes respectively. The images are distributed throughout the United States, captured under diverse conditions reflecting different sensor characteristics, shadow conditions, scene conditions and temporal attributes giving rise to large within-class variations. A few example images are shown in figure 4.

B. Setup

We evaluated our approach on the above datasets using individual and combined feature descriptors. We measure the

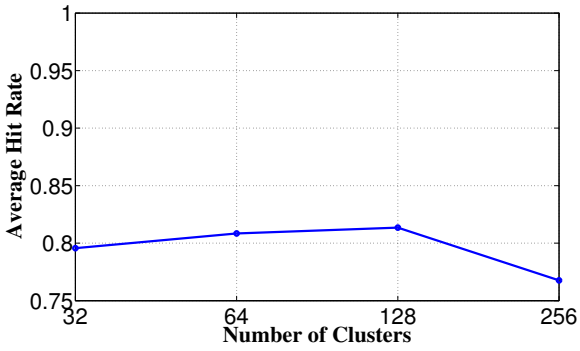

Fig. 5. Hit Rate versus cluster size evaluated on *MobileHomePark* dataset.

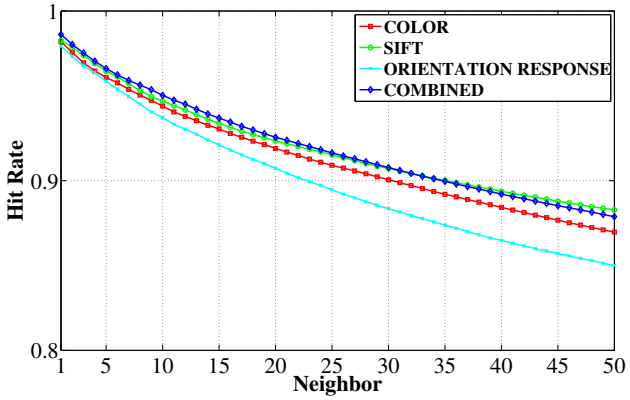

Fig. 6. Categorization accuracy for the top 50 matches in the *MobileHomePark* dataset.

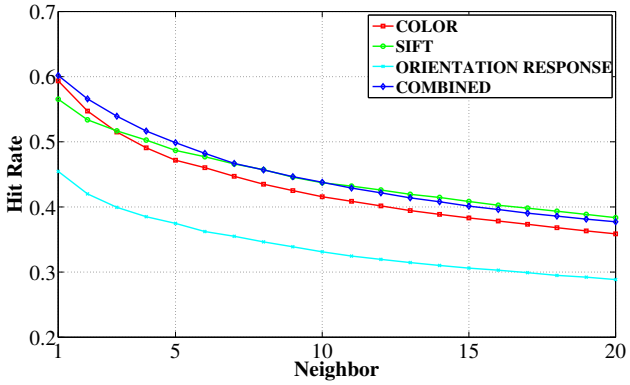

Fig. 7. Categorization accuracy for the top 20 matches in the *UCMERCED* dataset.

performance by measuring the *hit-rate* which is defined as the ratio of the number of retrieved images belonging to the same scene category as the query image to the total number of images in the scene category. We set parameters $d = 10$, $q = 16$, and $s = 8$. To set parameter m, we repeated the experiment with different values of m to determine the optimal *hit-rate* measures. We chose $m = 64$ as the computational gains with smaller m significantly outperforms the accuracy advantage obtained for $m = 128$. We evaluated our approach using the three feature descriptors. Additionally, we combined the different feature descriptors into a single vector to evaluated the performance. The distance between the query image and the rest of the images in the dataset is computed based on equation (1) and the 20 closest images are examined to measure the *hit-rate*. We compute *hit-rate* for all the images in the dataset and report the average *hit-rate*.

V. Results

Figure 6 shows the performance on the *MobileHomePark* dataset. As seen from the plot, all the four feature representations have similar accuracy for top 20 matches with combined having a slight edge over the rest. For the top 50 matches, the SIFT and the combined feature representations produce the best performance with average *hit-rate* 0.9197 and 0.9203 respectively. Figure 11 shows a few example query results for the *MobileHomePark* dataset.

Next, we assess the performance of the *UCMERCED* dataset. Figure 7 shows the average *hit-rate* obtained on this dataset. Except for the oriented responses, all the feature representation produced similar performance . We also examined the average *hit-rate* across different categories. Figure 8 shows the average *hit-rate* per individual scene categories. Our experiments show that scene with recurring patterns such as the parking lots, harbors produced high *hit-rate* whereas scene categories such as the airplane, storage tank, and tennis courts produced low *hit-rate*. The main reason could be that these scene categories require additional shape based feature representation to effectively characterize the scene. Figure 12 shows example query results from the *UCMERCED* dataset.

Finally, we examine the performance of the proposed approach on the *ORNL-I* dataset. Figure 9 shows the average *hit-rate* for the dataset. The performance is similar to the previous dataset. As shown in figure 10 the performance of the proposed technique is relatively high for the suburban and

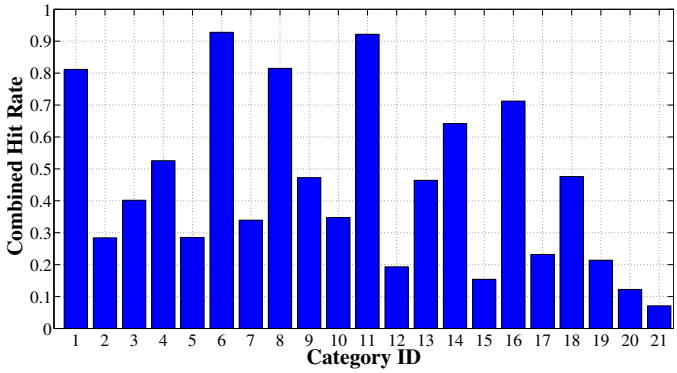

Fig. 8. Hit-rate for specific categories in the top 20 of the *UCMERCED* dataset. The categories in order from left to right: (1) *agricultural*, (2) *airplane*, (3) *baseball diamond*, (4) *beach*, (5) *buildings*, (6) *chaparral*, (7) *dense residential*, (8) *forest*, (9) *freeway*, (10) *golf course*, (11) *harbor*, (12) *intersection*, (13) *medium residential*, (14) *mobile home park*, (15) *overpass*, (16) *parking lot*, (17) *river*, (18) *runway*, (19) *sparse residential*, (20) *storage tanks*, and (21) *tennis courts*. Example queries from select categories (*harbor*, *parking lot*, and *storage tank*) can be seen in Figure 12.

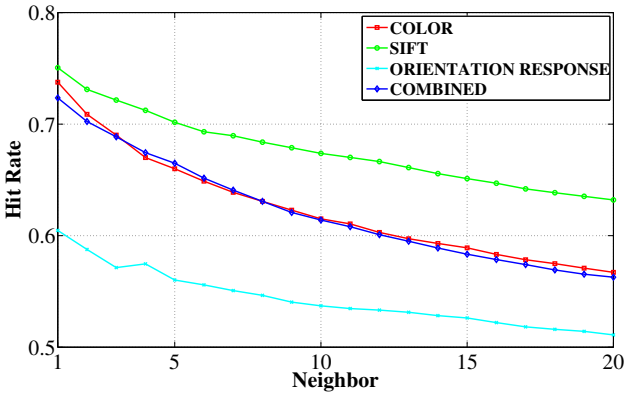

Fig. 9. Categorization accuracy for the top 20 matches in the *ORNL-I* dataset.

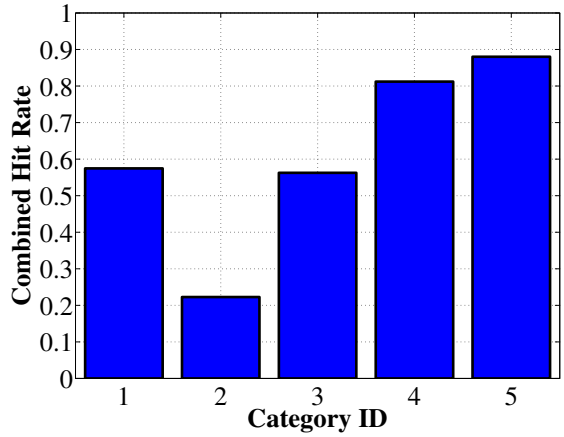

Fig. 10. Hit-rate for specific categories in the top 20 of the *ORNL-I* dataset.The categories on order from left to right: (1) *agricultural*, (2) *large-facility*, (3) *commercial*, (4) *suburban*, and (5) *wooded* . Example queries from select categories (*large-facility*, *commercial*, and *suburban*) can be seen in Figure 13.

downtown scene categories. The high within-class variations resulting from the differences in sensor characteristics, shadow conditions, scene conditions and temporal attributes impose limitations on the performance of the proposed approach. Figure 13 shows example query results from the *UCMERCED* dataset.

VI. Conclusion

In this paper, we have presented a scene representation technique that accounts for the spatial arrangement of low-level features. We evaluated the method with different feature descriptors. Our results indicate the spatial correlgoram based on SIFT features yield promising results. The performance can be further improved by incorporating additional features to characterize specific object level shape features. Additionally, the scalability of the method and the feasibility of the approach scene classification and change detection needs to be further explored.

Acknowledgment

This research was supported by the U.S. Department of Energy (DOE) Office of Science through the Summer Undergraduate Laboratory Internship (SULI) program. This manuscript has been authored by employees of UT-Battelle, LLC, under contract DE-AC05-00OR22725 with the U.S. Department of Energy. Accordingly, the United States Government retains and the publisher, by accepting the article for publication, acknowledges that the United States Government retains a non-exclusive, paid-up, irrevocable, world-wide license to publish or reproduce the published form of this manuscript, or allow others to do so, for United States Government purposes.

References

[1] M. Pesaresi and A. Gerhardinger, "Improved textural built-up presence index for automatic recognition of human settlements in arid regions with scattered vegetation," *Selected Topics in Applied Earth Observations and Remote Sensing, IEEE Journal of*, vol. 4, no. 1, pp. 16–26, march 2011.

[2] I. A. Rizvi and B. K. Mohan, "Object-based image analysis of high-resolution satellite images using modified cloud basis function neural network and probabilistic relaxation labeling process," *IEEE Trans. on Geoscience and Remote Sensing*, vol. 49, no. 12, pp. 4815–4820, 2011.

[3] R. Bellens, S. Gautama, L. Martinez-Fonte, W. Philips, J. C.-W. Chan, and F. Canters, "Improved classification of VHR images of urban areas using directional morphological profiles," *IEEE Trans. on Geoscience and Remote Sensing*, vol. 46, no. 10, pp. 2803–2813, 2008.

[4] A. K. Shackelford and C. H. Davis, "A combined fuzzy pixel-based and object-based approach for classification of high-resolution multispectral data over urban areas," *IEEE Trans. on Geoscience and Remote Sensing*, vol. 41, no. 10, pp. 2354–2363, 2003.

[5] P. Gamba, F. Dell'Acqua, G. Lisini, and G. Trianni, "Improved VHR urban area mapping exploiting object boundaries," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 45, no. 8, pp. 2676–2682, Aug. 2007.

[6] J. Huang, S. Kumar, M. Mitra, W. Zhu, and R. Zabih, "Spatial color indexing and applications," *International Journal of Computer Vision*, vol. 35, no. 3, pp. 245–268, 1999.

[7] L. Bruzzone and L. Carlin, "A multilevel context-based system for classification of very high spatial resolution images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 44, no. 9, pp. 2587–2600, sept. 2006.

[8] C. Unsalan and K. L. Boyer, "Classifying land development in high-resolution panchromatic satellite images using straight-line statistics," *IEEE Trans. on Geoscience and Remote Sensing*, vol. 42, pp. 907–919, 2004.

[9] X. Huang, L. Zhang, and P. Li, "Classification and extraction of spatial features in urban areas using high-resolution multispectral imagery," *IEEE Geoscience and Remote Sensing Letters*, vol. 4, no. 2, pp. 260–264, april 2007.

[10] D. G. Lowe, "Object recognition from local scale-invariant features," in *Proceedings of the International Conference on Computer Vision, Corfu*, 1999.

[11] T. Leung and J. Malik, "Representing and recognizing the visual appearance of materials using three-dimensional textons," *International Journal of Computer Vision*, vol. 43, no. 1, pp. 29–44, june 2001.

[12] Y. Yang and S. Newsam, "Bag-of-visual-words and spatial extensions for land-use classification," in *Proceedings of the ACM International Conference on Advances in Geographic Information Systems*, 2010.

[13] V. Vijayaraj, A. Cheriyadat, P. Sallee, B. Colder, R. R. Vatsavai, E. A. Bright, and B. L. Bhaduri, "Overhead image statistics," in *IEEE Applied Imagery Pattern Recognition Workshop*, 2008.

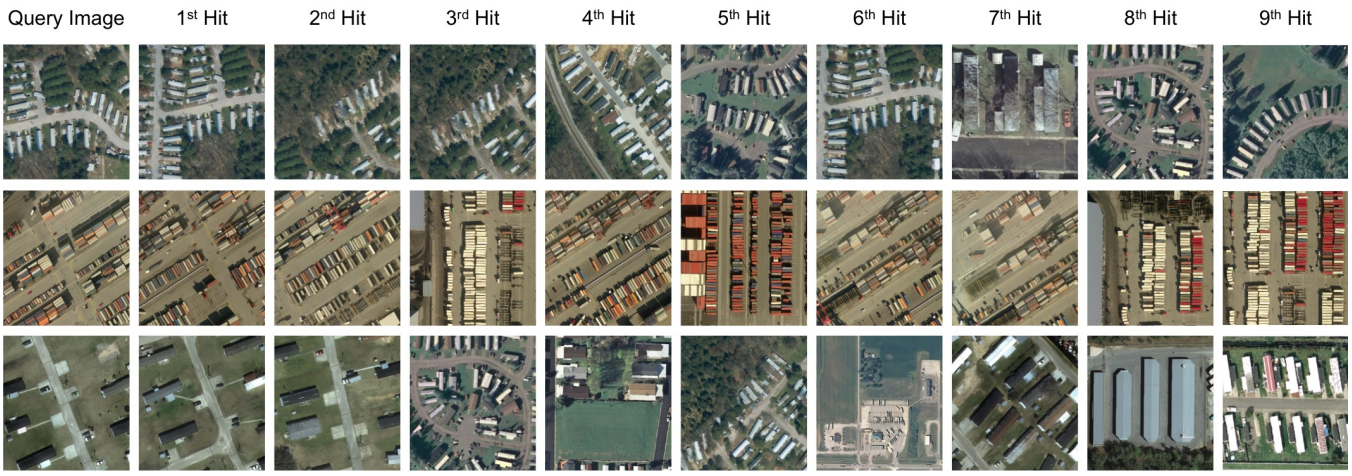

Fig. 11. Example Queries in the *MobileHomePark* dataset. Mobile home park category query on the top and bottom row. Other structures category query on the middle row

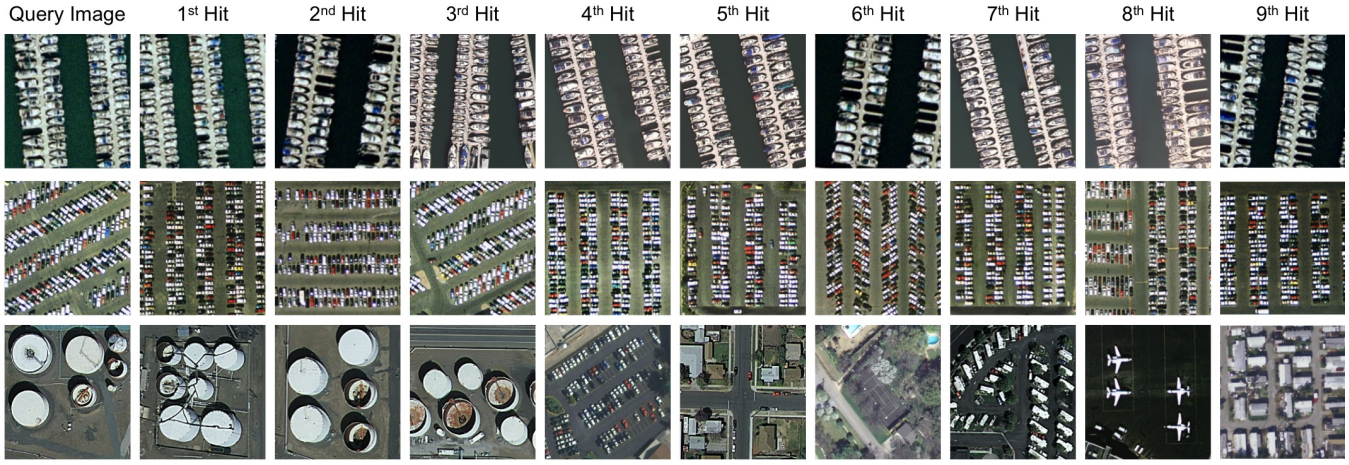

Fig. 12. Example Queries in the *UCMERCED* dataset. *Harbor* category query on the top row. *Parking lot* category query on the middle row. *Storage tank* category query on the bottom row.

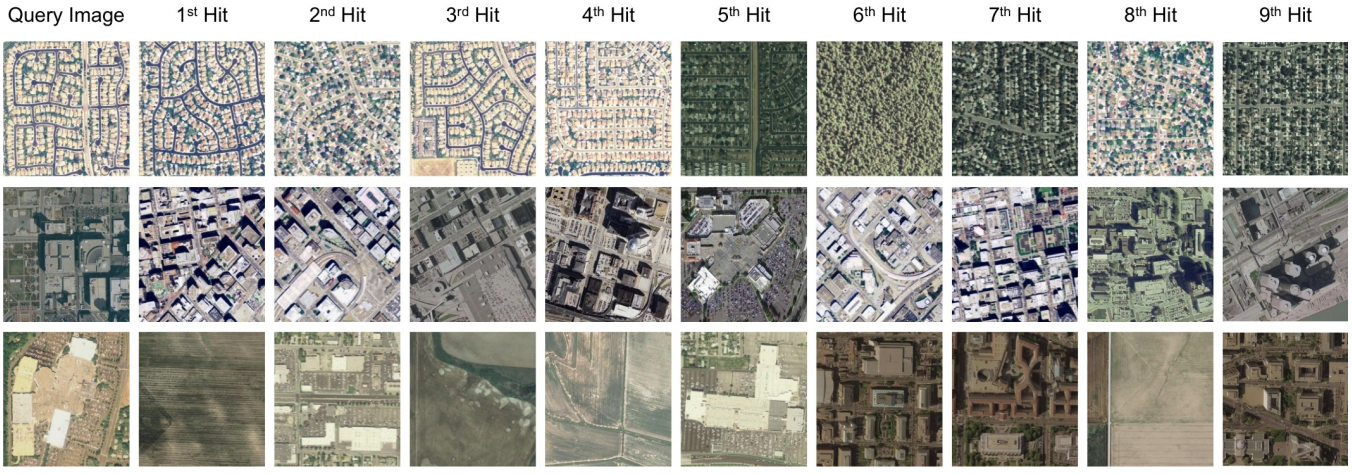

Fig. 13. Example Queries in the *ORNL-I* dataset. *Suburban* category query on the top row. *Commerical* category query on the middle row. *Large-facility* category query on the bottom row.