Aerial Image Classification Using Structural Texture Similarity

Vladimir Risojević and Zdenka Babić
Faculty of Electrical Engineering, University of Banja Luka
Patre 5, 78000 Banja Luka, Bosnia and Herzegovina
vlado@etfbl.net, zdenka@etfbl.net

Abstract*—*There is an increasing need for algorithms for automatic analysis of remote sensing images and in this paper we address the problem of semantic classification of aerial images. For the task at hand we propose and evaluate local structural texture descriptor and similarity measure. Nearest neighbor classifier based on the proposed descriptor and similarity measure, as well as image-to-class similarity, improves classification rates over the state-of-the-art on two datasets of aerial images. We evaluate the design choices and show that rich subband statistics, perceptually-based structural texture similarity measure and image-to-class similarity all contribute to the good performance of our classifier.

Keywords*—*Aerial image classification, Structural texture similarity, Local image descriptors

I. INTRODUCTION

One of the most important problems in aerial image analysis is semantic classification. The ultimate goal of semantic classification of aerial images is to assign a class from a predefined set, e.g. urban, industry, forest, etc., to each image pixel. Since aerial images are frequently multispectral and of high resolution, in order to reduce computational complexity, this problem is usually approached by dividing the aerial image into tiles, and assigning a class from a predefined set to each tile. Thus obtained classification of image tiles can then be used in content-based image retrieval or for constructing a thematic map, for example.

There has been a long history of using computer vision techniques for classification of aerial images. Some efforts are part of content-based aerial image retrieval systems, e.g. in [1] the authors use Gabor descriptors and self-organizing maps for classification of aerial images in order to enable efficient content-based retrieval from the database of aerial images. Parulekar et al. [2] classify satellite images into four semantic categories in order to enable fast and accurate browsing of the image database.

In [3] target detection in aerial images is performed. However, images are only classified into two classes, based on whether they contain the target object or not. Multi-class classification was investigated in [4] where aerial images are classified based on color, texture and structure features. In a more recent work [5], Ozdemir and Aksoy use bag-of-words model and frequent subgraph mining to construct higher level features for satellite image classification.

Various image descriptors and classification algorithms for aerial image classification are also actively evaluated. In [6] SIFT descriptors and Gabor texture descriptors are compared, as well as Gist descriptors and Gabor texture descriptors in [7]. In [8] a large-scale evaluation of eight various texture and color descriptors was performed on a new, currently the largest publicly available, dataset of aerial images.

In this paper we evaluate structural texture similarity measure, originally proposed in [9] and further developed in [10], on the task of aerial image classification. To the best of our knowledge, this is the first large scale evaluation of structural texture similarity measure on the task of aerial image classification. We also propose modifications which make structural texture similarity better suited for aerial image classification.

Structural texture similarity compares texture images using 7 × 7-pixels sliding windows at corresponding locations. Comparing only windows at corresponding locations limits the spatial layout of images. To avoid this drawback, in Section II, we regard images as unordered sets of windows, and allow for comparing of windows at different spatial locations. We accomplish this by introducing local image descriptors, which use larger windows at more sparsely distributed locations.

At the moment, learning-based, parametric classifiers predominate aerial image classification tasks [4], [5], [6], [7], [8]. However, nearest neighbor (NN) classifier has several advantages over learning-based classifiers: (i) allows a large number of classes, as well as easy adding of new classes and labeled examples, (ii) requires no training phase, and (iii) has no problems with overfitting. Inspired by the results from [11] we chose nearest neighbor classifier based on image-to-class similarity, as described in Section III.

We perform evaluations on two publicly available datasets, already used in the literature on the subject [7], [8]. In Section IV we give details about the performed experiments and compare the classification accuracies with the results from the literature. It is worth noting that the dataset used in [8] is the largest dataset that has been used for aerial image classification to date. We show that, for both datasets, our results improve state-of-the-art.

II. LOCAL STRUCTURAL TEXTURE SIMILARITY DESCRIPTOR

Structural texture similarity was originally proposed in [9] and further developed in [10]. In order to compute structural texture similarity for two texture images, each image is first convolved by a complex steerable filter bank and for each 7 × 7-pixels sliding window in each subband the following subband-coefficient statistics are computed: subband means, standard deviations, autocorrelation coefficients for one pixel displacements along horizontal and vertical axes, as well as cross-correlations between subbands. The similarity of these two textures is then computed by comparing the subband statistics on corresponding sliding windows, i.e. the windows at the same locations in images.

A drawback of this approach is its sensitivity to changes in the spatial layouts of images. This is useful in natural scene classification where the spatial layout is a strong cue for classification [12]. On the other hand, it was shown that absolute spatial layout taken into account via spatial pyramid kernel [8] and Gist descriptors [7] does not improve classification accuracy for remote sensed images. Consequently, we do not consider the absolute spatial layout as a useful cue for remote sensed image classification.

In order to avoid these drawbacks and still make use of the descriptive ability of structural texture representation we decided to derive local image descriptors from it. Local image descriptors are very popular in scene classification and object recognition [13]. Some popular local descriptors were also evaluated in the context of remote sensed image classification in [6] and [8]. Therefore, we compute descriptors for windows on a regular grid and consider an image as a bag of descriptors, without paying attention to their absolute spatial location. Compared to the originally proposed descriptor, we dropped the autocorrelation terms used in [10] because of their sensitivity to rotation, we use larger window sizes, and we do not restrict the window size to 7 × 7 pixels. To the best of our knowledge this is the first attempt to use structural texture descriptors as local image descriptors.

Computation of structural texture descriptors starts with convolving the input image with a complex steerable filter bank at S scales and K orientations, resulting in SK subbands. Let $W_k^{\mathbf{x}}$ denotes the coefficients of the subband $k = 1, \ldots, SK$, in the window $\mathbf{x}$. For each window $\mathbf{x}$ the following statistics are computed:

- Means of subband coefficient magnitudes
$$\mu_k^{\mathbf{x}} = E\left\{|W_k^{\mathbf{x}}|\right\} \tag{1}$$
- Standard deviations of subband coefficient magnitudes
$$\sigma_k^{\mathbf{x}} = E\left\{\left(|W_k^{\mathbf{x}}| - \mu_k^{\mathbf{x}}\right)^2\right\} \tag{2}$$
- Cross-correlation coefficients between magnitudes of coefficients in subbands k and l, at the same scale, but different orientations, as well as at the same orientation but different scales
$$\rho_{kl}^{\mathbf{x}} = \frac{E\left\{\left(|W_k^{\mathbf{x}}| - \mu_k^{\mathbf{x}}\right)\left(|W_l^{\mathbf{x}}| - \mu_l^{\mathbf{x}}\right)\right\}}{\sigma_k^{\mathbf{x}}\sigma_l^{\mathbf{x}}},\ k \neq l. \tag{3}$$

In the above equations the expected values are computed as empirical averages over subband windows.

These statistics are collected in a local descriptor. In the remainder of this paper, we will refer to this descriptor as *structural texture (ST)* descriptor. ST descriptor is related to Gabor texture descriptor [14], which uses means and standard deviations of subbands globally, and to Gist descriptor [12], which uses means of subband blocks on a 4×4 grid.

Structural texture similarity (STSIM) between two windows, $\mathbf{x}$ and $\mathbf{y}$, represented by their ST descriptors, is computed by averaging their similarities in all subbands,

$$Q(\mathbf{x}, \mathbf{y}) = \frac{1}{SK}\sum_{k=1}^{SK} l_k^{\frac{1}{3}}(\mathbf{x}, \mathbf{y})\, c_k^{\frac{1}{3}}(\mathbf{x}, \mathbf{y})\, r_k^{\frac{1}{3}}(\mathbf{x}, \mathbf{y}), \tag{4}$$

where $l_k(\mathbf{x}, \mathbf{y})$ is the similarity of mean values

$$l_k(\mathbf{x}, \mathbf{y}) = \frac{2\mu_k^{\mathbf{x}}\mu_k^{\mathbf{y}} + K_1}{\left(\mu_k^{\mathbf{x}}\right)^2 + \left(\mu_k^{\mathbf{y}}\right)^2 + K_1}, \tag{5}$$

and $c_k(\mathbf{x}, \mathbf{y})$ is the similarity of standard deviations

$$c_k(\mathbf{x}, \mathbf{y}) = \frac{2\sigma_k^{\mathbf{x}}\sigma_k^{\mathbf{y}} + K_2}{\left(\sigma_k^{\mathbf{x}}\right)^2 + \left(\sigma_k^{\mathbf{y}}\right)^2 + K_2}. \tag{6}$$

In equations (5) and (6) K_1 and K_2 are small constants introduced for stability reasons.

The term $r_k(\mathbf{x}, \mathbf{y})$ is the contribution of cross-correlation coefficients and it is obtained by averaging similarities of all cross-correlation coefficients as follows:

$$r_k(\mathbf{x}, \mathbf{y}) = \frac{1}{N_\rho}\left[\sum_{k \neq l}\left(1 - 0.5|\rho_{kl}^{\mathbf{x}} - \rho_{kl}^{\mathbf{y}}|\right)\right], \tag{7}$$

where $N_\rho = S + K - 2$ is the number of cross-correlation coefficients for a subband block.

III. NEAREST NEIGHBOR STRUCTURAL TEXTURE SIMILARITY CLASSIFIER

The problem of image classification, in general, consists of assigning a test image $\mathbf{X}$ to a class from the predefined set $\{C_1, C_2, \ldots, C_k\}$. Traditionally, a nearest neighbor classifier involves computation of similarities between the test image and labeled images (image-to-image similarity) and classifies the test image into the class containing the most similar labeled image. In this paper, inspired by [11], rather than computing image-to-image similarities we take a different approach, and compute the image-to-class similarities between the test image and each of the classes. We classify the test image to the class for which this similarity measure attains its maximum.

We consider an image to be a set of windows or blocks $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$, which are represented using its ST-descriptors, as described. If an image $\mathbf{X}$ belongs to a class C, we consider that all its blocks $\mathbf{x}_i$ belong to the same class and write $\mathbf{x}_i \in C$. We define block-to-class similarity as

$$Q(\mathbf{x}, C) = \max_{\mathbf{y} \in C} Q(\mathbf{x}, \mathbf{y}). \tag{8}$$

Similarity of the test image $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ to the class C is $Q(\mathbf{X}, C) = \prod_{i=1}^{n} Q(\mathbf{x}_i, C)$. The computation of image-to-class similarity is illustrated in Fig. 1. For each block from the test image the most similar block in the set of labeled images is found. The overall image-to-class similarity is the product of individual block-to-class similarities for each of the blocks from the test image.

Now, classification of the test image $\mathbf{X}$ proceeds in the following way:

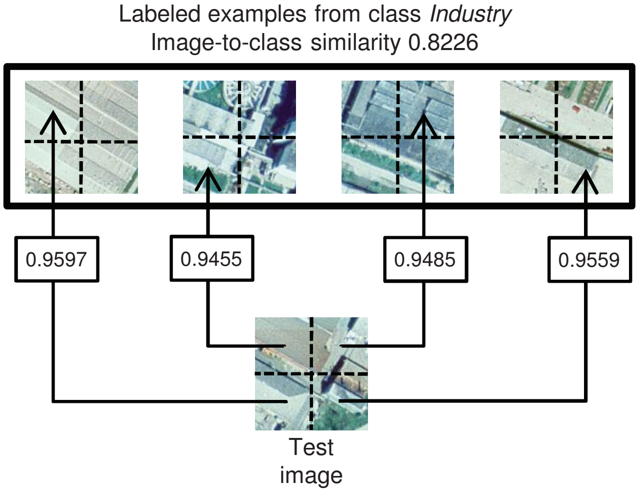

Labeled examples from class *Industry*
Image-to-class similarity 0.8226
0.9597
0.9455
0.9485
0.9559
Test
image

Fig. 1. Image-to-class similarity. Block-to-class similarities are indicated on the arrows. See text. (Best viewed in color.)

1) Calculate ST-descriptors for blocks $\mathbf{x}_1, \ldots, \mathbf{x}_n$ of the test image $\mathbf{X}$;
2) For each block $\mathbf{x}_i, i = 1, \ldots, n$ of the test image and for each class C_j, $j = 1, \ldots, k$ compute the similarity $Q(\mathbf{x}_i, C_j)$ using (8);
3) Compute image-to-class similarities between the test image and each class $Q(\mathbf{X}, C_j) = \prod_{i=1}^{n} Q(\mathbf{x}_i, C_j)$, $j = 1, \ldots, k$;
4) Classify $\mathbf{X}$ to the class $\hat{C} = \arg\max_C Q(\mathbf{X}, C)$.

The advantage of using image-to-class similarity over image-to-image similarity for image classification can be observed in Fig. 2, where image-to-image similarities of a test image to a set of labeled examples from two classes are shown. The most similar to the test image is the labeled image belonging to the negative class, so if image-to-image similarity had been used, the test image would have been classified incorrectly. On the other hand, when image-to-class similarity is used, the test image will be classified correctly, since image-to-class similarity for the positive class is larger than for the negative one.

IV. EXPERIMENTAL RESULTS

In order to evaluate the proposed descriptors and classifier, we conduct experiments on two datasets of remote sensed images: Banja Luka dataset and UC Merced dataset. These datasets have recently been used in similar experiments in [7], [8]. We found that in both cases our classifier advances the state of the art. Most notably, our classifier outperforms all the other approaches on UC Merced dataset [8], which is the largest publicly available dataset of remote sensed images.

A. Banja Luka Dataset

For our first experiment we used Banja Luka dataset[1], which consists of 606 images of size 128 × 128 pixels. The images were manually classified into 6 classes: houses, cemetery, industry, field, river, and trees.

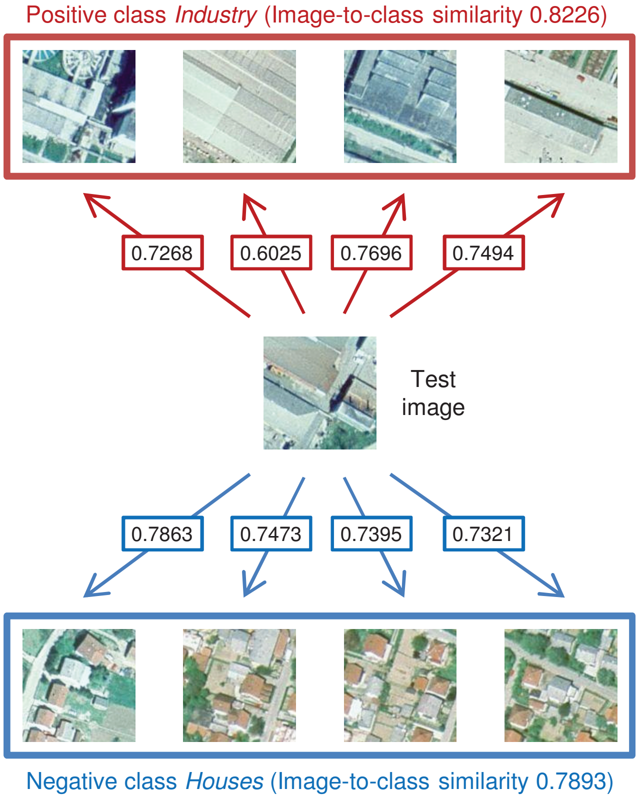

Positive class *Industry* (Image-to-class similarity 0.8226)
0.7268
0.6025
0.7696
0.7494
Test
image
0.7863
0.7473
0.7395
0.7321
Negative class *Houses* (Image-to-class similarity 0.7893)

Fig. 2. Image-to-image similarity versus image-to-class similarity. Image-to-image similarities between the test image and the labeled images are given on the arrows. Using image-to-image similarity the test image would have been classified incorrectly, but using image-to-class similarity the classification is correct. (Best viewed in color.)

In the feature extraction phase, we decompose each image using both complex steerable filter bank and Gabor filter bank at 4 scales and 6 orientations. We included Gabor filter bank into this analysis motivated by its biological plausibility and good performance of Gabor-based descriptors in [6] and [7]. Next we compute ST-descriptors for subband blocks on 1×1 (which corresponds to global subband coefficients statistics), 2×2, and 4×4 grids. For color images we compute the descriptors and similarities for individual color-components of the RGB colorspace and average the similarities.

Since the distribution of images in the classes is very uneven, we use half of the images from each class as labeled images, and the other half as test images. We repeat the experiment 10 times with different random splits of the dataset, and average the results.

First we would like to investigate the influence of the design choices on the performance of the nearest neighbor structural texture similarity based (NN-STSIM) classifier. We compare NN-STSIM classifiers based on complex steerable filters with (a) NN-STSIM classifiers based on Gabor filters, as well as NN-classifiers using: (b) L_1-norm induced metric with ST-descriptors and image-to-class similarity, (c)

[1] Available: http://dsp.etfbl.net/aerial/

image-to-image similarity with STSIM, and (d) STSIM with means and standard deviations of subband blocks only. For this evaluation we use only grayscale versions of images. The comparisons are made for different numbers of subband blocks, and the performances are shown in Fig. 3. We can see that the classifiers using ST-descriptors based on Gabor filters (labeled with Gabor in Fig. 3) outperform the classifiers based on complex steerable filters (labeled with steerable). For this reason in the subsequent experiments we use only Gabor filter bank. Compared to other classifier variants, we can see that the classifiers using STSIM have consistently better performance than those using L_1-norm induced metric, which is due to the perceptual basis of STSIM. Furthermore, classifier benefits from richer feature statistics and the classifier with full descriptors is always better than the classifier using only means and standard deviations of subband blocks. Finally, we can see that image-to-class classifier outperforms image-to-image classifier, which is consistent with our toy example in Fig. 2 and findings of [11].

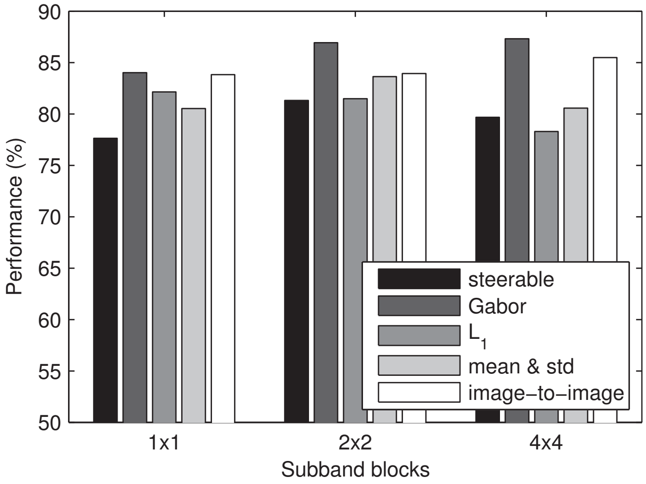


Fig. 3. Comparison of performances of various ST-descriptor based classifiers.

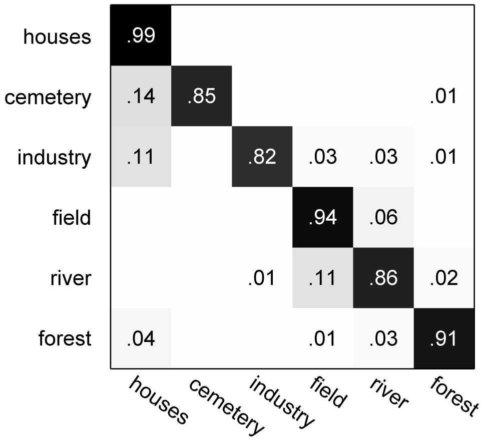
We also compared our classification accuracies to the results obtained in [7] using SVM classifier with Gabor and Gist descriptors, and the results are given in Table I. In the table, our classifier is labeled NN-STSIM with indicated subband partition. In the last two rows are the results from [7].

Table I. Classification accuracies for Banja Luka dataset.

Classifier	**Grayscale (%)**	**Color (%)**
NN-STSIM 1×1	84.0	86.8
NN-STSIM 2×2	86.9	**89.6**
NN-STSIM 4×4	**87.3**	88.3
SVM Gabor [7]	84.5	88.0
SVM Gist [7]	79.5	89.3

We can see that, for grayscale images, NN-STSIM 4×4 classifier has the best performance and it even slightly outperforms SVM-based approaches. For color images, we obtained the best performance with NN-STSIM 2×2 classifier, and again, it slightly outperforms SVMs.

Confusion matrix for color NN-STSIM 2×2 classifier is given in Fig. 4. We note that confusions mainly arise between classes which can be difficult even for humans. The most notable examples are houses and industry versus cemetery, because of rectangular structures with strong oriented edges, and river versus field, because both have homogeneous, smooth texture without pronounced edges. It is also important to note that there are not many confusions between natural (river, trees, field) and man-made classes (houses, cemetery, industry).



	houses	cemetery	industry	field	river	forest
houses	.99					
cemetery	.14	.85				.01
industry	.11		.82	.03	.03	.01
field				.94	.06	
river			.01	.11	.86	.02
forest	.04			.01	.03	.91

Fig. 4. Confusion matrix for Banja Luka dataset.

B. UC Merced Dataset

UC Merced dataset[2] consists of aerial images of 21 land-use classes. All images are 256×256 pixels and in RGB colorspace. They are manually classified into 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. Each class contains 100 images, which makes this dataset the largest publicly available dataset for remote sensed image classification.

In [8] the authors compared eight popular local and global image descriptors with SVM classifiers on this dataset. Somewhat surprisingly, color histogram in HLS colorspace outperformed the other approaches, and global and local texture-based approaches using only intensity information were pretty much leveled.

For this dataset, we compute ST-descriptors in the same way as for Banja Luka dataset, except that we use only Gabor filter bank. Following the protocol in [8] we performed five-fold cross-validation, with ${}^{4}/_{5}$ of the dataset used as labeled images and ${}^{1}/_{5}$ as test images. The obtained performances

[2] Available: http://vision.ucmerced.edu/datasets

Table II. Classification accuracies for UC Merced dataset.

Descriptor	Grayscale (%)	Color (%)
NN-STSIM 1×1	77.2	80.0
NN-STSIM 2×2	80.9	81.4
NN-STSIM 4×4	**83.4**	**86.0**
Gabor [8]	76.9	80.5
Bag-of-words [8]	76.8	N/A
BoW + SCK [8]	77.7	N/A
HLS histogram [8]	N/A	81.2

are shown in Table II along with selected results from [8], namely: Gabor texture descriptor, bag-of-words, bag-of-words with spatial co-occurrence kernel (BoW + SCK) and color histogram in HLS colorspace. We can see that NN-STSIM 2×2 and NN-STSIM 4×4 classifiers for grayscale images outperform all the other approaches that use only intensity information. Moreover, the performance of NN-STSIM 4×4 classifier for grayscale images is better than that of HLS histogram. From these results, it is obvious that our classifier better utilizes intensity information than the state-of-the-art approaches do.

Classification accuracy is further improved by including color information, and NN-STSIM classifiers for color images consistently outperform the NN-STSIM classifiers for grayscale images. The accuracy obtained with NN-STSIM 4×4 for color images is the best overall with correct classification rate of 86%.

In Fig. 5, per-class classification rates for NN-STSIM 4×4 classifier are shown. The classes on which we obtained the highest classification rates are: beach, chaparral, harbor, and runway. The images in these classes have pretty much homogeneous texture. The lowest classification rates are obtained on buildings and intersection, and to a certain extent on storage tanks and tennis courts. These are the classes with large intra-class variations.

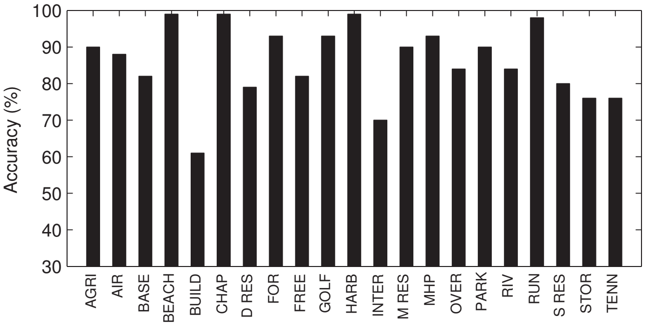


Fig. 5. Per-class classification rates for UC Merced dataset.

Confusion matrix for NN-STSIM 4×4 classifier for color images is shown in Fig. 6. The most notable confusions are between freeway and overpass, dense residential and medium density residential, as well as between building and storage tanks. Visual similarity of these classes is high even for human observers.

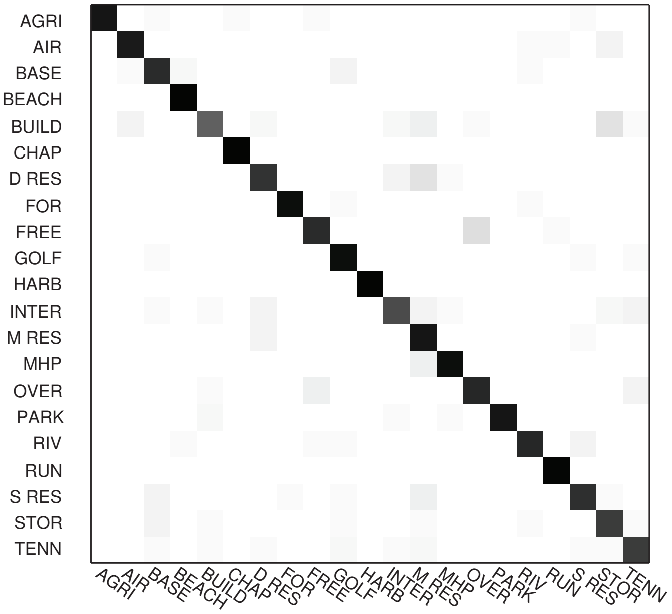


Fig. 6. Confusion matrix for UC Merced dataset.

V. Conclusion

The main contribution of this paper is the evaluation of local structural texture descriptors and structural texture similarity with nearest neighbor classifier for semantic classification of aerial images. Our experiments on two datasets show that NN classifiers with perceptually-based local descriptors and similarity measure, as well as image-to-class similarity have better performance than learning-based approaches for aerial image classification. Most notably, our classifier improves state-of-the-art on the 21-class dataset used in [8], which has been the largest dataset for aerial image classification to date. It is also interesting to note that increasing the number of categories and images causes much less degradation of classification accuracy for NN classifier compared to learning-based classifiers.

Besides evaluation we also proposed using structural texture descriptor as local image descriptor in a bag-of-descriptors fashion. We evaluated a number of design choices and showed that using rich statistics of subband coefficients is beneficial to classifier performance. Furthermore, since structural texture similarity metric is perceptually-based it shows better performance than often used L_1-norm induced similarity metric. Finally, our classifier uses image-to-class similarity which outperforms image-to-image similarity, traditionally used in nearest-neighbor classifiers.

The main drawback of our classifier is its computational complexity. Traditional nearest neighbor classifiers using L_p-norm induced metrics are made practical by means of indexing structures, e.g. kd-trees. However, at the moment there are no indexing structures appropriate for structural similarity metric, and it is an attractive direction of future research. Another possibility is to use a high-performance platform, such as GPGPU, for the computation of STSIM.

References

[1] W.-Y. Ma and B. S. Manjunath, “A texture thesaurus for browsing large aerial photographs,” *J. Am. Soc. Inform. Sci.*, vol. 49, no. 7, pp. 633–648, May 1998.

[2] A. Parulekar, R. Datta, J. Li, and J. Z. Wang, “Large-scale satellite image browsing using automatic semantic categorization and content-based retrieval,” in *IEEE Int. Workshop on Semantic Knowledge in Computer Vision*, 2005, pp. 1873–1880.

[3] Z. Li and L. Itti, “Saliency and gist features for target detection in satellite images,” *IEEE Trans. Image Proc.*, vol. 20, no. 7, pp. 2017–2029, July 2011.

[4] J. Fauqueur, N. G. Kingsbury, and R. Anderson, “Semantic discriminant mapping for classification and browsing of remote sensing textures and objects,” in *Proc. ICIP*, 2005, pp. 846–849.

[5] B. Ozdemir and S. Aksoy, “Image classification using subgraph histogram representation,” in *Proc. ICPR*, 2010, pp. 1112 – 1115.

[6] Y. Yang and S. Newsam, “Comparing SIFT descriptors and Gabor texture features for classification of remote sensed imagery,” in *Proc. ICIP*, October 2008, pp. 1852 –1855.

[7] V. Risojević, S. Momić, and Z. Babić, “Gabor descriptors for aerial image classification,” in *Proc. ICANNGA, Part II*, vol. 6594 of *LNCS*, pp. 51–60. Springer Berlin / Heidelberg, 2011.

[8] Y. Yang and S. Newsam, “Bag-of-visual-words and spatial extensions for land-use classification,” in *Proc. ACM SIGSPATIAL GIS*, 2010, pp. 270–279.

[9] X. Zhao, M.G. Reyes, T.N. Pappas, and D.L. Neuhoff, “Structural texture similarity metrics for retrieval applications,” in *Proc. ICIP*, 2008, pp. 1196–1199.

[10] J. Zujovic, T.N. Pappas, and D.L. Neuhoff, “Structural similarity metrics for texture analysis and retrieval,” in *Proc. ICIP*, 2009, pp. 2225 –2228.

[11] O. Boiman, E. Shechtman, and M. Irani, “In defense of nearest-neighbor based image classification,” in *Proc. CVPR*, 2008, pp. 1–8.

[12] A. Oliva and A. Torralba, “Modeling the shape of the scene: a holistic representation of the spatial envelope,” *Int. J. Comp. Vis.*, vol. 42, no. 3, pp. 145–175, 2001.

[13] J. Zhang, M. Marszalek, S. Lazebnik, and C. Schmid, “Local features and kernels for classification of texture and object categories: A comprehensive study,” *Int. J. Comp. Vis.*, vol. 73, pp. 213–238, June 2007.

[14] B. S. Manjunath and W.-Y. Ma, “Texture features for browsing and retrieval of image data,” *IEEE Trans. on Pat. Anal. and Mach. Intel.*, vol. 18, no. 8, pp. 837–842, August 1996.