Fusion of Global and Local Descriptors for Remote Sensing Image Classification

Vladimir Risojević, *Student Member, IEEE,* Zdenka Babić, *Member, IEEE*

***Abstract*—Very high resolution remote sensing images offer increased amount of details available for image interpretation. However, despite enhanced resolution these details result in spectral inhomogeneities, making automated image classification more difficult. In this letter we propose to combine texture and local image features to address this problem. We first address the Enhanced Gabor Texture Descriptor which is a global descriptor based on cross-correlations between subbands, and show that it achieves very good results in classification of aerial images showing a single thematic class. Next, the performances obtained on individual land cover/land use classes using our global texture descriptor and local SIFT descriptor are compared. We identify classes of images best suited for each descriptor, and argue that these descriptors encode complementary information. Finally, a hierarchical approach for the fusion of global and local descriptors is proposed and evaluated over a number of classifiers. The proposed descriptor fusion approach exhibits significantly improved classification results, reaching the accuracy of around 90%.**

***Index Terms*—Remote sensing image classification, Gabor texture descriptor, SIFT descriptor, stacked generalization**

I. Introduction

REMOTE sensing image classification is an active research topic spurred by the need to analyze continuously growing body of remote sensing imagery. When the problem at hand involves classification of very high resolution images we cannot rely on spectral homogeneity any more, and must use other elements of image interpretation, such as texture, shape, pattern, size, etc. In an attempt to endow image classification algorithms with this capability, object-oriented approach to remote sensing image classification has emerged [1].

In this letter we are concerned with tile-based classification of aerial images into land cover/land use classes and make several contributions. First we propose a novel texture descriptor based on cross-correlations between spatial-frequency subbands of Gabor image decomposition, which are ignored in the original Gabor texture descriptor [2]. We named that descriptor *Enhanced Gabor Texture Descriptor (EGTD)*. Next, we compare per-class performances of both EGTD and bag-of-words representations and identify classes on which one of the descriptors outperforms the other. Finally, since these representations are complementary, we propose a method for fusion of the information obtained using both global and local image descriptors. We experimentally show that this method significantly improves the classification performance.

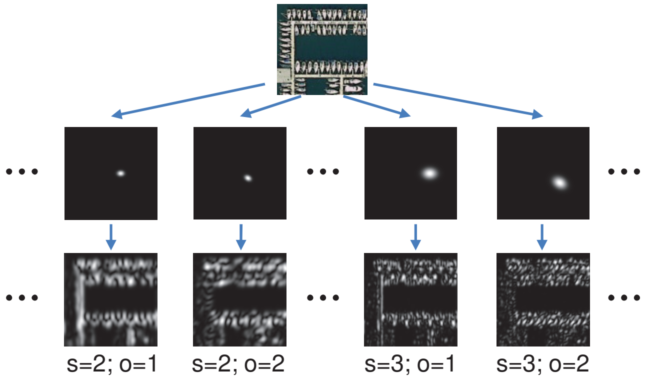

Fig. 1. Examples of amplitude responses of Gabor filters (middle row) and magnitudes of Gabor coefficients (bottom row) at different scales (s) and orientations (o) for an aerial image.

One of the most popular texture descriptors in remote sensing applications is Gabor texture descriptor [2]. In a recent study [3] it was shown that Gabor texture descriptors yield reasonable performance on large datasets of texture images in the absence of affine and non-affine transformations in the spectral domain. In remote sensing image analysis, it has been used for image retrieval [4] and classification [5], [6]. Recently, its extension to hyperspectral images has been proposed [7].

Orthogonal subband transforms used in image coding aim at removing the correlations from image representation. It has been noted [8], however, that the magnitudes of wavelet coefficients in different subbands are correlated. This can be observed in Fig. 1, where magnitudes of Gabor coefficients at several scales and orientations are shown. The magnitudes of coefficients tend to have similar values at the same relative spatial locations in subbands. This fact has been used in texture synthesis [9], as well as in texture similarity assessment for image compression [10]. Also in [11], a local image descriptor using cross-correlations between magnitudes of Gabor coefficients at different scales or orientations has been proposed for aerial image classification. It has been shown that such a descriptor outperforms the original Gabor texture descriptor.

Gabor wavelets are not orthogonal and there are also correlations between raw coefficients at different scales and/or orientations. Correlations between Gabor coefficients at different scales and the same orientation have been shown to correspond to center-surround organization (opponency) of the cells in human retina and used for texture [12] and satellite image classification [13], as well as target detection in satellite

Manuscript received July 17, 2012; revised September 14, 2012; accepted October 12, 2012

This work was supported in part by the Ministry of the Science and Technology of the Republic of Srpska under contract 06/0-020/961-220/11

The authors are with the Faculty of Electrical Engineering, University of Banja Luka, Patre 5, 78000 Banja Luka, Bosnia and Herzegovina (e-mail: vlado@etfbl.net; zdenka@etfbl.net)

images [14]. In [15], orientation difference descriptor for aerial image classification which uses cross-correlations between coefficients at different orientations and at the same scale has been proposed. In this letter we propose enhanced Gabor texture descriptor (EGTD) which aggregates means and standard deviations of Gabor coefficients as well as cross-correlations between coefficients at different scales or orientations.

The proposed EGTD is mainly texture oriented. However, there are land use classes which are entirely defined by individual objects present in images, e.g. storage tanks, baseball fields, intersections, etc. Due to the averaging of the wavelet coefficients over the image domain, EGTD is unable to encode local information. In that case local descriptors heavily used in general object recognition could give better results.

One of the most popular local descriptors at the moment is Scale Invariant Feature Transform (SIFT) proposed in [16]. It is mainly used within bag-of-words (BoW) framework which includes vector quantization of descriptors and a pooling step in order to estimate their probability distribution in an image. BoW classifiers of aerial images have been thoroughly investigated in [5] and [17].

EGTD and SIFT encode complementary information about the image and their fusion could improve the classification performance. Our approach for descriptor fusion is hierarchical. We first train classifiers for both descriptors. Then we concatenate the confidence scores returned by the used classifiers and obtain the mid-level representation. Mid-level descriptors are then used as inputs to the classifier at the second level (metalearner). Recently, similar schemes, using support vector machines as metalearners, have been proposed for image classification, both general [18] and remote sensing [19]. These hierarchical classifiers are, in fact, using *stacked generalization* or *stacking* [20] as a way to combine classifiers. It has been reported in the machine learning literature that even with very simple metalearners, such as linear [21] or regularized linear regression [22], good performance can be achieved. In this letter we evaluate various metalearners at the task of remote sensing image classification, and show that most of them yield similar classification accuracies which makes simple classifiers very appealing choices for metalearners.

The rest of the paper is organized in the following way. In Section II, we present the theory behind the EGTD descriptor. A review of stacking is given in Section III. Experimental setup and results are reported in Section IV.

II. Enhanced Gabor Texture Descriptor

The starting point for the construction of the enhanced Gabor texture descriptor is a Gabor filter bank at S scales and K orientations. The impulse responses of the filters are scaled and rotated versions of the Gabor function

$$g\left(x,y\right)=\frac{1}{2\pi\sigma_x\sigma_y}\exp\left[-\frac{1}{2}\left(\frac{x^2}{\sigma_x^2}+\frac{y^2}{\sigma_y^2}\right)+j\omega x\right]. \quad (1)$$

Suppose we have an image $I\left(x,y\right),\left(x,y\right)\in\Omega$, where Ω is the set of image pixels. The output of the Gabor filter with impulse response $g_{mn}\left(x,y\right)$ at scale $m=1,\ldots,S$ and orientation $n=1,\ldots,K$, is given by the convolution

$$W_{mn}\left(x,y\right)=I\left(x,y\right)*g_{mn}\left(x,y\right). \quad (2)$$

In [2] Gabor texture descriptor which consists of means (energies) and standard deviations of the modules of filter responses at all scales and orientations has been proposed

$$\mu_{mn}=\iint\limits_{\Omega}\left|W_{mn}\left(x,y\right)\right|dxdy\ , \quad (3)$$

$$\sigma_{mn}=\sqrt{\iint\limits_{\Omega}\left(\left|W_{mn}\left(x,y\right)\right|-\mu_{mn}\right)^2dxdy}\ . \quad (4)$$

Let $W_{mn}\left(x,y\right)$ and $W_{m'n}\left(x,y\right)$ be the responses of Gabor filters (2) at orientation n and scales m and m'. Based on [12], we define the opponent features $\psi_{mm'n}$, as the energies of the differences of normalized filter responses

$$\Delta W_{mm'n}\left(x,y\right)=\left|\frac{W_{mn}\left(x,y\right)}{\mu_{mn}}-\frac{W_{m'n}\left(x,y\right)}{\mu_{m'n}}\right|, \quad (5)$$

$$\psi_{mm'n}=\iint\limits_{\Omega}\Delta W_{mm'n}\left(x,y\right)dxdy. \quad (6)$$

Note that our definition of the opponent features is slightly different from the one in [12] because in the original paper real-valued Gabor filters were used. The opponent features can be regarded as measures of cross-correlations between the filter responses at different scales.

The responses of the filters at the same scale but different orientations can also be correlated depending on the type of land cover. To the best of our knowledge those correlations have not been used for image analysis thus far. Analogously to (6) we propose using the energies of the differences of the normalized filter responses at the same scale m and different orientations, n and n', as an image descriptor

$$\Delta W_{mnn'}\left(x,y\right)=\left|\frac{W_{mn}\left(x,y\right)}{\mu_{mn}}-\frac{W_{mn'}\left(x,y\right)}{\mu_{mn'}}\right|, \quad (7)$$

$$\rho_{mnn'}=\iint\limits_{\Omega}\Delta W_{mnn'}\left(x,y\right)dxdy. \quad (8)$$

We also add standard deviations of the differences (5) and (7) to the descriptor

$$\upsilon_{mm'n}=\sqrt{\iint\limits_{\Omega}\left|\Delta W_{mm'n}\left(x,y\right)-\psi_{mm'n}\right|^2dxdy}, \quad (9)$$

$$\nu_{mnn'}=\sqrt{\iint\limits_{\Omega}\left|\Delta W_{mnn'}\left(x,y\right)-\rho_{mnn'}\right|^2dxdy}. \quad (10)$$

Finally, the EGTD is built by aggregating the quantities given by equations (3), (4), (6), (8), (9) and (10) into a $SK\left(S+K\right)$-dimensional vector $\mathbf{f}=\left[f_1,\ldots,f_{SK(S+K)}\right]^T$.

For computing the dissimilarity between images we adopt distance metric based on the weighted L_1-norm

$$\text{dist}\left(\mathbf{f}^{(a)},\mathbf{f}^{(b)}\right)=\sum_{j=1}^{SK(S+K)}\left|\frac{f_j^{(a)}-f_j^{(b)}}{\alpha\left(f_j\right)}\right|, \quad (11)$$

where $\mathbf{f}^{(a)}$ and $\mathbf{f}^{(b)}$ are descriptors of two single-band images, and $\alpha\left(f_j\right)$ are the standard deviations of the respective features over the training set.

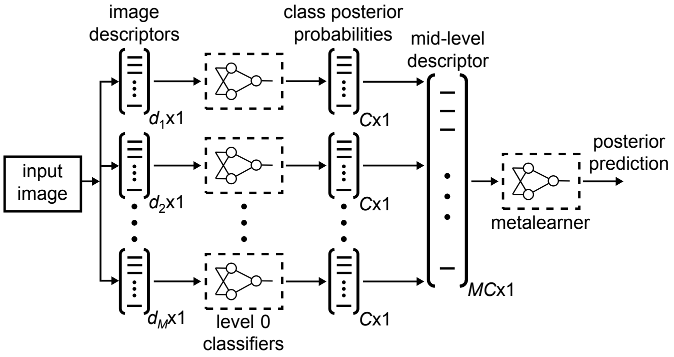

Fig. 2. Block diagram of the hierarchical descriptor fusion scheme.

III. Fusion of Descriptors

In this letter we use stacking for fusing the descriptors. Our stacking scheme uses two levels of classifiers as shown in Fig. 2. The classifiers at the first level (level-0) are used to build the mid-level image representation. It is then fed to the level-1 classifier (metalearner) which outputs the final prediction.

Let us suppose that we have a set of C classes into which the images have to be classified. The training set contains N labelled images. For each image, M descriptors are computed, $\mathbf{f}_i = [f_{i1}, f_{i2}, \dots, f_{id_i}]^T$, where $i = 1, \dots, M$ and d_i is the dimensionality of the i-th descriptor.

We train M level-0 classifiers, one for each descriptor, and their outputs, $\mathbf{p'}_i = [p'_{i1}, p'_{i2}, \dots, p'_{iC}]^T = cl(\mathbf{f}_i)$, $i = 1, \dots, M$, are confidence scores that the input image should be classified into each of C classes. They are transformed into posterior probabilities using the softmax function

$$p_{ij} = \frac{\exp\left(\beta p'_{ij}\right)}{\sum_{k=1}^{C} \exp\left(\beta p'_{ik}\right)}, i = 1, \dots, M, j = 1, \dots, C, \quad (12)$$

where β controls the "sharpness" of the function. The posterior probabilities are concatenated into a mid-level descriptor $\mathbf{P} = [\mathbf{p}_1, \dots, \mathbf{p}_M]^T$. The dimensionality of this descriptor is MC. Finally, the metalearner is trained using the mid-level descriptors. The metalearner outputs the posterior prediction $\hat{c} = ml(\mathbf{P})$.

There is an interesting variant of the described scheme, proposed in [23]. Instead of using all confidence scores for training the level-1 classifier, we can train C classifiers, one for each class, using only confidence scores for that class, $z_c = ml_c(p_{1c}, p_{2c}, \dots, p_{Mc})$, $c = 1, \dots, C$. The test instance is then classified to the class $\hat{c} = \arg\max_{c=1,\dots,C} z_c$. This approach is known as *stackingC*.

In order to prevent overfitting, the mid-level representation is obtained using a procedure reminiscent of cross-validation [20]. The training set is randomly split into P parts. One of the parts is held out as validation set and level-0 classifiers are trained using the rest of the training set. Then mid-level representation is built for the samples from the validation set. The procedure is repeated taking each of the P parts as the validation set in turn. Thus we obtain mid-level representations for all the samples from the training set and the metalearner is trained using the complete training set.

IV. Experiments

We perform the experiments on a dataset of aerial images from the UC Merced used in [17]. All images are RGB, 256×256 pixels, with pixel resolution of one foot (0.3 meters). They are manually classified into 21 classes, corresponding to various land cover and land use types: 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.

In each experiment we randomly split the dataset into the training and test sets. The training set is used to train the level-0 classifiers, using $P = 5$, and to train the metalearner, as described in Section III. The performances of the trained classifiers are then assessed using the test set which was not used for training the classifier. We repeat each experiment on five different random training/test splits of the dataset and report means and standard deviations of the obtained classification accuracies. We report results obtained using 10, 50 and 90 training samples per class.

EGTD is computed using a Gabor filter bank at 4 scales and 6 orientations. This results in 240-dimensional descriptors. For classification we tested SVMs with linear, χ^2, radial basis function (RBF) and generalized RBF kernel using metric (11),

$$K\left(\mathbf{f}^{(a)}, \mathbf{f}^{(b)}\right) = \exp\left[-\operatorname{dist}\left(\mathbf{f}^{(a)}, \mathbf{f}^{(b)}\right)\right]. \quad (13)$$

As a local descriptor, standard bag-of-words (BoW) descriptor is chosen. It is obtained by computing SIFT descriptors on a regular grid and vector quantizing them using a codebook with 1000 codewords. Histogram of codeword occurrences is a 1000-dimensional BoW image descriptor. In the study [24], χ^2 kernel is shown to perform best in this case. However, after reviewers' comments we tested the same kernels as for EGTD.

Multi-class classification is obtained by training one-vs-all SVMs for all classes and classifying a test sample to the class which corresponds to the maximal SVM response.

A. Global vs. Local Descriptors

We are first concerned with the performances of the tested kernels for EGTD and BoW descriptors. From the results in Table I we can see that for both descriptors the generalized RBF kernel is a suitable choice.

Comparing the EGTD and BoW-based classifiers, we see that the EGTD-based classifier slightly outperforms the BoW-based one. The good performance of the EGTD is due to suitability of textural features for classification of remote sensing images and its ability to represent features salient over scales and/or orientations. Besides the overall classification performances of these two descriptors, we are also interested in their performances on individual classes. In order to better assess discriminative abilities of descriptors we train the classifiers using 10 training samples per class. Such a small number of training samples accentuates the need for good image representations because there could exist significant variations in images. On the other hand, had we used larger

TABLE I
CLASSIFICATION ACCURACIES (%) OF GLOBAL AND LOCAL DESCRIPTORS.

		Number of training images per class		
Descr.	**Kernel**	10	50	90
EGTD	linear	48.03 ± 3.07	67.49 ± 0.42	73.71 ± 1.92
	χ^2	58.88 ± 2.83	76.76 ± 1.17	83.33 ± 2.50
	RBF	61.16 ± 2.90	81.20 ± 1.14	86.76 ± 2.19
	(13)	**65.31 ± 2.22**	**82.88 ± 0.85**	**88.19 ± 1.56**
BoW	linear	53.42 ± 2.83	71.98 ± 1.00	77.33 ± 1.64
	χ^2	**62.86 ± 1.70**	79.28 ± 0.87	84.29 ± 2.21
	RBF	57.53 ± 2.05	76.72 ± 1.02	83.52 ± 2.79
	(13)	62.32 ± 2.11	**79.77 ± 0.71**	**84.29 ± 1.30**

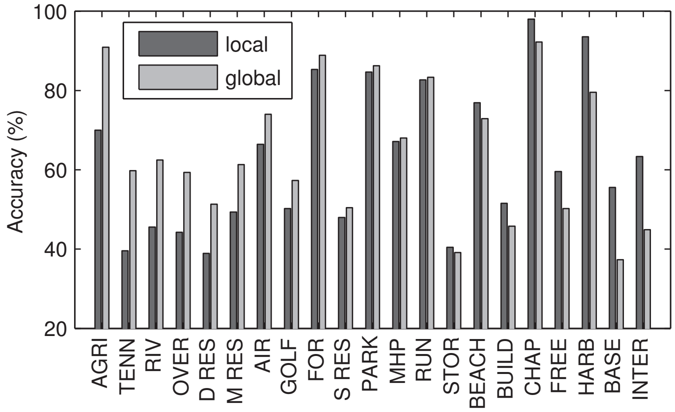

Fig. 3. Comparison of global and local image descriptors. Cases when a particular descriptor outperforms the other can be clearly seen.

number of training samples, the results for some classes would have reflected the dataset bias instead of the discriminative abilities of the descriptors.

In Fig. 3 the breakdown of classification accuracies accross classes for both global and local descriptors is given. On the left side of the bar graph are the classes on which global descriptor outperforms the local one. Going to the right the performance of the classifier using local descriptors improves and, finally, on the right side are the classes on which the local descriptor is better. We can see that the global descriptor is better on classes such as *agricultural, tennis court* and *river*, which contain image-scale features and are mainly texture oriented, Fig. 4 top row. On the other hand, the local descriptor is better on classes which contain distinctive structures whose (lack of) presence is used to classify the images, e.g. *harbor, baseball diamond* and *intersection* as can be seen in the bottom row of Fig. 4.

These results suggest that global and local image descriptors contain complementary information and their fusion should improve the performance of the classifier. Simple concatenation of the descriptors would result in a new descriptor of high dimensionality, which increases the chance of overfitting. Therefore, in the following we investigate stacking as an approach for building ensembles of image classifiers.

B. Hierarchical Classifier

In order to combine the cues from both the global and local descriptors we build a mid-level representation as described in Section III. We compute the posterior probabilities using (12) with $\beta = 1$, and normalize the obtained vectors to zero mean and standard deviation of unity for each input sample. Finally, mid-level representation is obtained by concatenating these normalized vectors.

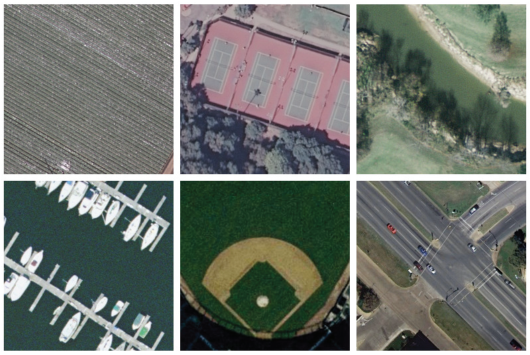

Fig. 4. Images from the classes on which better results are obtained using global descriptor (top row) and local descriptor (bottom row).

TABLE II
CLASSIFICATION ACCURACIES (%) OF VARIOUS METALEARNERS AND THEIR STACKINGC VARIANTS (MARKED WITH C) FOR FUSION OF GLOBAL AND LOCAL DESCRIPTORS.

	Number of training images per class		
Metalearner	10	50	90
Concatenation	60.69 ± 2.17	81.03 ± 0.96	86.67 ± 2.54
Linear SVM	66.39 ± 0.92	88.46 ± 0.92	92.38 ± 1.54
Linear SVM C	71.04 ± 3.73	88.90 ± 0.88	92.38 ± 1.87
Linear regr.	68.16 ± 3.08	88.65 ± 0.75	92.67 ± 1.53
Linear regr. C	70.16 ± 3.54	88.46 ± 0.90	91.81 ± 1.79
Ridge regr.	70.69 ± 3.25	88.70 ± 0.92	92.67 ± 1.29
Ridge regr. C	70.93 ± 1.36	88.36 ± 0.78	92.48 ± 1.48
RBF SVM	71.02 ± 3.02	**89.43 ± 0.79**	93.05 ± 1.56
RBF SVM C	71.06 ± 2.99	88.50 ± 1.03	93.05 ± 2.06
RBF ridge regr.	68.62 ± 2.80	88.93 ± 0.76	**93.90 ± 1.09**
RBF ridge regr. C	**71.41 ± 2.84**	88.65 ± 0.96	92.95 ± 2.27

Thus obtained mid-level representation is then fed into a metalearner. We evaluate the following metalearners: linear regression, ridge regression, linear SVM, RBF SVM and ridge regression with RBF kernel, as well as their stackingC variants. Lasso regression had not brought any improvements so we decided to leave it out. Multi-class classification is performed in the same way as for level-0 classifiers.

The obtained results are given in Table II. The accuracies for simple concatenation of image descriptors are given in the first row. It is the simplest way of fusion of descriptors and in this case no metalearner is actually used. However, its performance is consistently worse compared to the cases when metalearners are used. As for tested metalearners we can see that the achieved accuracies are very similar. From these results we make several observations:

- There is no need to use kernels in metalearners. Linear classifiers do equally well with reduced complexity. This conclusion is intuitively appealing because the mid-level representation consists of confidence scores of level-0 classifiers. Linear metalearners combine these confidence scores giving higher weights to outputs of those level-0 classifiers having high confidence of predicting the correct class. On the other hand, classifiers with RBF kernels strongly depend on Mahalanobis distance between

TABLE III
CLASSIFICATION ACCURACIES (%) AFTER DIMENSIONALITY REDUCTION OF EGTD.

	Number of training images per class		
Metalearner	10	50	90
RDA EGTD	56.72 ± 3.38	75.58 ± 2.03	81.90 ± 2.33
Linear SVM C	**67.44 ± 2.21**	86.57 ± 1.20	90.48 ± 2.05
RBF SVM	66.84 ± 2.23	**87.56 ± 1.28**	**91.05 ± 2.55**

the samples. However, it is not clear how Mahalanobis distance in the feature space of confidence scores should be interpreted.

- Since the dimensionality of the mid-level representation is relatively low, there is no need to use SVM whatsoever. In this particular case, the dimensionality of the mid-level descriptors is $2C = 42$, and it can be seen that regression-based classifiers perform very well.
- The results for stackingC suggest that the metalearners can be trained using mid-level descriptors whose dimensionality equals the number of different descriptors, two in this case. Regression-based classifiers with two-dimensional descriptors have very low complexity.

C. Dimensionality reduction

The dimensionality of EGTD can be fairly high, e.g. it is 240 in the described experiments. Therefore, we explored dimensionality reduction of EGTD using principal component analysis, linear discriminant analysis, and regularized linear discriminant analysis (RDA). We obtained the best results applying RDA separately to three parts of EGTD, namely means and standard deviations of subbands, means and standard deviations of subband differences at the same orientation and different scales, and means and standard deviations of subband differences at the same scale and different orientations. The dimensionality of the resulting descriptor is $3 \times 20 = 60$. For classification at level-0 we used SVM with the same kernel as for the original descriptor. In Table III are given the results for level-0 classifier (RDA EGTD) and for the two best metalearners. The results for other metalearners are similar. Although the accuracies when using only the level-0 classifier are reduced 6-8%, the accuracies of the resulting hierarchical classifiers are at most 4% lower. Therefore, we believe that the dimensionality reduction in this case has interesting potential and should be investigated further.

V. CONCLUSION

In this letter we proposed EGTD and stacking-based approach for very high resolution remote sensing image classification. We experimentally showed that EGTD and BoW representations are complementary and their fusion significantly improves the classification performance. Descriptor fusion using stackingC with linear or ridge regression performs very well in terms of effectiveness and complexity.

It is important to note that this method for descriptor fusion is general. It is not restricted to the particular descriptors and classifiers used in this letter, so it can be used with different descriptors and classifiers, provided that the classification confidence scores for classes are available.

In the future work we plan to investigate fusion of more descriptors and, particularly, the inclusion of color information.

REFERENCES

[1] J. B. Campbell, *Introduction to remote sensing*. Guilford Press, 2006.

[2] B. S. Manjunath and W.-Y. Ma, “Texture features for browsing and retrieval of image data,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 18, pp. 837–842, Aug. 1996.

[3] U. Kandaswamy, S. Schuckers, and D. Adjeroh, “Comparison of texture analysis schemes under nonideal conditions,” *IEEE Trans. Image Proc.*, vol. 20, pp. 2260–2275, Aug. 2011.

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

[5] 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.*, Oct. 2008, pp. 1852–1855.

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

[7] T. Bau, S. Sarkar, and G. Healey, “Hyperspectral region classification using a three-dimensional gabor filterbank,” *IEEE Trans. Geosci. Remote Sens.*, vol. 48, pp. 3457–3464, Sep. 2010.

[8] E. P. Simoncelli, “Statistical modeling of photographic images,” in *Handbook of Image and Video Processing*, A. C. Bovik, Ed. Academic Press, 2005, pp. 431–441.

[9] J. Portilla and E. P. Simoncelli, “A parametric texture model based on joint statistics of complex wavelet coefficients,” *Int. J. Comput. Vision*, vol. 40, no. 1, pp. 49–71, Oct. 2000.

[10] J. Zujovic, T. Pappas, and D. Neuhoff, “Structural similarity metrics for texture analysis and retrieval,” in *Proc. IEEE Int. Conf. Image Process.*, 2009, pp. 2225–2228.

[11] V. Risojević and Z. Babić, “Aerial image classification using structural texture similarity,” in *Proc. IEEE Int. Symp. Signal Process. and Inform. Technology*, Bilbao, Spain, Dec. 2011, pp. 190–195.

[12] A. Jain and G. Healey, “A multiscale representation including opponent color features for texture recognition,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 7, pp. 124–128, Jan. 1998.

[13] D. Dai and W. Yang, “Satellite image classification via two-layer sparse coding with biased image representation,” *IEEE Geosci. Remote Sens. Lett.*, vol. 8, pp. 173–176, Jan. 2011.

[14] Z. Li and L. Itti, “Saliency and gist features for target detection in satellite images,” *IEEE Trans. Image Proc.*, vol. 20, pp. 2017–2029, Jul. 2011.

[15] V. Risojević and Z. Babić, “Orientation difference descriptor for aerial image classification,” in *Proc. 19th Int. Conf. Systems, Signals and Image Process. (IWSSIP)*, Vienna, Austria, Apr. 2012, pp. 156–159.

[16] D. Lowe, “Distinctive image features from scale-invariant keypoints,” *Int. J. Comput. Vision*, vol. 60, no. 2, pp. 91–110, November 2004.

[17] 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.

[18] A. Abdullah, R. C. Veltkamp, and M. A. Wiering, “Spatial pyramids and two-layer stacking SVM classifiers for image categorization: A comparative study,” in *Proc. of Int. Joint Conf. Neural Networks*, Jun. 2009, pp. 5–12.

[19] G. Sheng, W. Yang, L. Chen, and H. Sun, “Satellite image classification using sparse codes of multiple features,” in *Proc. 10th IEEE Int. Conf. Signal Process.*, Oct. 2010, pp. 952–955.

[20] D. H. Wolpert, “Stacked generalization,” *Neural Networks*, vol. 5, pp. 241–259, 1992.

[21] K. M. Ting and I. H. Witten, “Issues in stacked generalization,” *J. of Artificial Intelligence Research*, vol. 10, pp. 271–289, 1999.

[22] S. Reid and G. Grudic, “Regularized linear models in stacked generalization,” in *Proc. 8th Int. Workshop Multiple Classifier Syst.*, ser. MCS ’09. Berlin, Heidelberg: Springer-Verlag, 2009, pp. 112–121.

[23] A. K. Seewald, “How to make stacking better and faster while also taking care of an unknown weakness,” in *Proc. 19th Int. Conf. Mach. Learning*, Sydney, Australia, Jul. 2002, pp. 554–561.

[24] 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.