Benchmarking of Remote Sensing Segmentation Methods

Stanislav Mikeš, Michal Haindl, *Senior Member, IEEE*, Giuseppe Scarpa, and Raffaele Gaetano

***Abstract*—We present the enrichment of the Prague Texture Segmentation Data-Generator and Benchmark (PTSDB) to include the assessment of the remote sensing (RS) image segmenters. The PTSDB tool is a Web-based (http://mosaic.utia.cas.cz) service designed for real-time performance evaluation, mutual comparison, and ranking of various supervised or unsupervised static or dynamic image segmenters. PTSDB supports rapid verification and development of new segmentation approaches. The RS datasets contain ten spectral Advanced Land Imager (ALI) satellite images, their RGB subsets, and very-high-resolution GeoEye RGB images, with optional additive-noise-resistance checking. Alternative setting options allow us to also test scale, rotation, or illumination invariance. The meaningfulness of the newly proposed dataset is demonstrated by testing and comparing several RS segmentation algorithms, and showing that the benchmark figures provide a solid framework for the fair and critical comparison among different techniques.**

***Index Terms*—Benchmark, remote sensing (RS) segmentation, supervised segmentation, unsupervised segmentation.**

I. Introduction

SATELLITE image segmentation is currently a consolidated prerequisite for successful remote sensing (RS) scene analysis, used, e.g., in crop inventory, geological and environment surveys, and military applications. Recent advances in RS technologies, and the consequent increase in the availability of RS data, have further pushed forward the development of segmentation-based applications [1]–[4], and several commercial products in RS image analysis, such as the eCognition [5] and ENVI [6] suites, are by now equipped with sophisticated segmentation tools. This scenario motivates a growing number of research activities on image segmentation. However, the diversity in both the kind of available data sources and the targeted applications has given life to a vast range of approaches and solutions [7]–[11], which are not supported by reliable and objective means to compare the performance of different techniques.

Very limited efforts have been made, in fact, to develop suitable quantitative measures of segmentation quality, especially in the case of RS. In this field, it is, in fact, quite common that researchers use their own data and related ground-truths, which are not publicly available to others, and present only a few carefully selected positive examples as validation for a new algorithm. This habit definitely encourages the proposal of more and more new techniques, whatever their actual merits, rather than the advancement of the most promising image segmentation approaches.

The optimal approach to check several variants of a developed method by carefully comparing the results with the state-of-the-art in this area is practically impossible because most methods are either too complicated or insufficiently described to be implemented in an acceptable time frame. Since no benchmark oriented to the development of segmentation methods for RS is available, we have generalized the Prague Texture Segmentation Data-Generator and Benchmark (PTSDB) [12] for the RS data applications. The solution is implemented in the form of a Web-based data generator and benchmark software suite. In particular, our proposal is aimed at both facing the lack of a rich shared dataset for evaluation, and allowing for a deep critical view on each technique's advantages and drawbacks as well as fairer comparisons among different methods.

In fact, it is well known that proper testing and robust learning of performance characteristics require large test sets and objective ground-truths, which are unrealistic requirements for natural satellite images. Thus, the satellite test images that are actually used are inevitably few, and they share the same drawbacks—subjectively generated ground-truth regions and limited extent of such a set, which is very difficult and expensive to enlarge. These problems motivated our preference for random mosaics with randomly filled satellite textures even if they only approximately correspond to satellite scenes. The most appealing feature of this tradeoff is the unlimited number of different test images with the corresponding objective and free ground-truth map available for each of them.

Moreover, to cope with the diversity of methods and targeted applications, we opt for a feature-rich scheme, which relies, on one hand, on a precise algorithm taxonomy concerning the employment of user-provided information (supervised segmentation, number of classes, map selection from a hierarchical stack, etc.) and, on the other hand, on a wide spectrum of performance metrics. This latter aspect is particularly important, since it favors the emergence of qualifying points as well as their validation, through the analysis of the correspondences among related indicators.

Manuscript received December 05, 2013; revised June 20, 2014; accepted March 13, 2015. Date of publication April 19, 2015; date of current version July 20, 2015. This work was supported by the Czech Science Foundation under project GAČR 14-10911S.

S. Mikeš and M. Haindl are with the Institute of Information Theory and Automation, Czech Academy of Sciences, Prague 182 08, Czech Republic (e-mail: xaos,haindl@utia.cz).

G. Scarpa and R. Gaetano are with the University Federico II of Naples, Naples 80138, Italy (e-mail: giscarpa,raffaele.gaetano@unina.it).

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/JSTARS.2015.2416656

II. Benchmark

The PTSDB is a Web-based (http://mosaic.utia.cas.cz) service [12] designed for real-time performance evaluation, mutual comparison, and ranking of various supervised or unsupervised static or dynamic image segmenters. The key objective of the PTSDB is to compute several accuracy measures for each given algorithm over the selected dataset. Once different segmentations have been collected over a given dataset, it is then possible to score them with respect to any of the computed accuracy indicators. This is of critical importance for three main reasons as follows:

1) to check the progress of an algorithm's development;
2) to mutually compare any two methods;
3) to track and measure the progress toward human-level segmentation performance over time.

A correct experimental evaluation should compare the tested method to several leading alternative algorithms, using a sufficiently large test image dataset and employing several evaluation measures for such comparison (in the absence of one clearly superior measure). Contrary to the prevailing practice when single authors verify their methods on a few carefully selected and thus noninformative positive examples, our benchmark possesses all these mentioned important features. While the color benchmark textures were chosen intentionally to produce unusually difficult tests in order to leave large margins for better segmentation algorithms to be derived in the future, the ALI multispectral textures contain richer spectral information and their textural analysis thus is less demanding. The benchmark operates either in the full mode for registered users (unrestricted mode—U) or in a restricted mode. The benchmark allows users: to obtain customized experimental satellite texture mosaics and their corresponding ground-truths (U); to obtain the benchmark mosaic sets with their corresponding ground-truths; to evaluate working segmenters and compare them with the state-of-the-art methods; to update the benchmark database (U) with an algorithm's details; to assess robustness with respect to noise; to check single mosaics' evaluation details (the criteria values and the resulting thematic maps); to rank segmentation algorithms according to the most common benchmark criteria; to obtain LaTeX- or MATLAB-coded result tables (U); and to select a user-defined subset of the criteria (U).

A. RS Data

Generated texture mosaics as well as the benchmarks are composed of the following texture types: 1) gray-scale textures (derived from the corresponding color textures); 2) color textures; 3) bidirectional texture function (BTF) textures; 4) ALI and GeoEye multispectral satellite images; 5) dynamic textures; 6) rotation invariant texture sets; 7) scale-invariant texture sets; and 8) illumination invariant texture sets and several invariant combinations.

The RS benchmark proposed here uses the Advanced Land Imager (ALI) and the high-resolution GeoEye observations.

The EO-1 (Earth Observing-1—http://eo1.usgs.gov) ALI is the first Earth-observing instrument to be flown under NASA's New Millennium Program (NMP). The ALI employs novel wide-angle optics and a highly integrated multispectral and panchromatic spectrometer. The focal plane for this instrument is partially populated with four sensor chip assemblies (SCAs) and also covers $3^\circ \times 1.625^\circ$. Operating in a pushbroom fashion at an orbit of 705 km, the ALI provides Landsat-type panchromatic and multispectral bands. These bands have been designed to mimic six Landsat bands with three additional bands covering 0.433–0.453, 0.845–0.890, and 1.20–1.30 μm. The ALI also contains wide-angle optics designed to provide a continuous $15^\circ \times 1.625^\circ$ field of view for a fully populated focal plane with 30-m resolution for the multispectral pixels and 10-m resolution for the panchromatic pixels.

TABLE I
ALI and GeoEye Bands and Spectral Ranges

ALI	Band	Spectral range (μm)	Description
0000	(PAN)	0.048–0.69	Panchromatic
0001	(MS-1′)	0.433–0.453	VNIR(blue)
0002	(MS-1)	0.45–0.515	VNIR(blue)
0003	(MS-2)	0.525–0.605	VNIR(green)
0004	(MS-3)	0.63–0.69	VNIR(red)
0005	(MS-4)	0.775–0.805	VNIR
0006	(MS-4′)	0.845–0.89	VNIR
0007	(MS-5′)	1.2–1.3	SWIR
0008	(MS-5)	1.55–1.75	SWIR
0009	(MS-7)	2.08–2.35	SWIR

GeoEye	Band	Spectral range (μm)	Description
	(PS-1)	0.45–0.51	Blue
	(PS-2)	0.51–0.58	Green
	(PS-3)	0.655–0.69	Red

GeoEye-1 [13] was launched in 2008 and simultaneously captures image detail up to 0.41 m for panchromatic images and 1.65 m for multispectral images. The benchmark uses pan-sharpened (0.5-m resolution) color (RGB) images (fusion of multispectral and panchromatic bands).

ALI and GeoEye bands and spectral ranges are listed in Table I. The benchmark uses 31 multispectral ALI and 52 GeoEye color textures categorized into 12 thematic classes. The thematic classes on each satellite data differ. The satellite texture parts, which are not used in the corresponding test mosaics, are used as separate training sets in the benchmark-supervised mode.

B. Benchmark Sets Creation

Benchmark 512×512 test mosaics are built by means of a Voronoi polygon random generator, and filled with randomly selected ALI/GeoEye textures. It is worth emphasizing that smaller and irregularly shaped objects are more difficult to segment than larger and regular shaped (square or circular) ones. ALI/GeoEye benchmarks (multispectral and RGB) are generated upon request in three quantities (10, 40, and 90 test mosaics) either in unsupervised or supervised mode, the latter including additional separate training sets. If required, however, any number of such mosaics can be generated. With each texture mosaic, the corresponding ground-truth and mask images are included. The RS benchmark allows us to check the segmenter noise resistance. All generated mosaics may be corrupted with additive Gaussian, Poisson, or salt and pepper noise. Alternative benchmarks allow us to also test scale and rotation or illumination invariance of the evaluated segmentation algorithm.

C. Performance Evaluation

The uploaded benchmark segmentation results are assessed, (permanently—U) stored in the database, and used to rank the segmenter according to a chosen criterion. PTSDB uses the most common 27 evaluation criteria sorted into 4 thematic groups: 1) region-based (5+5) [14]; 2) pixel-wise (11+1); 3) consistency measures (2) [15]; and 4) clustering comparison criteria (3) [16]. The performance criteria mutually compare ground-truth image regions with the corresponding machine segmented regions. The basic region-based criteria available [14] are correct segmentation, over-segmentation, under-segmentation, missed error, and noise error. All these criteria are available either with a single threshold parameter setting or in the form of performance curves and their integrals. The pixel-wise group contains the most common classification criteria such as the omission and commission errors, class accuracy, recall, precision, and mapping score. The consistency criteria [15] are global and local consistency errors. Finally, the last set contains three clustering comparison measures [16]. A detailed description of all these criteria (see http://mosaic.utia.cas.cz [12]) would go beyond the scope of this paper.

Uploaded results are also grouped according to the level of user interaction of the corresponding segmentation technique. Three possibly concomitant flags can be set when uploading, specifying whether 1) the used method is supervised (training-based classification); 2) the uploaded result is hand-picked from a hierarchical stack; and 3) the number of different ground-truth objects is given *a priori*. The efficiency of segmentation methods themselves cannot be considered because the benchmark obtains final segmentation results only.

III. Sample Methods

A. A Binary Tree-Structured Segmenter Family

The dynamic hierarchical classifier (DHC) [17] is one of the best performing *unsupervised* algorithms tested on the ALI dataset. It follows a top-down binary splitting paradigm, providing a stack of nested hierarchical segmentations. In particular, it combines two different segmentation methods that share the same paradigm: 1) the tree-structured Markov random field (TS-MRF) algorithm [18] and 2) the recursive-TFR (R-TFR) [19], which is an evolution of the texture fragmentation and reconstruction (TFR) algorithm proposed in [20]. The former is a spectral-oriented method where spatial regularity is controlled by means of an MRF [21] prior model. The latter is a texture-oriented method especially suited for *macro*-textured images. Given the DHC methodological roots, the authors have primarily used the benchmark to assess its performances w.r.t. TS-MRF and R-TFR, which had been previously tested on the same ALI dataset. These three intimately interrelated techniques are briefly recalled in the following.

TS-MRF is a recursive segmentation algorithm. The whole image of interest, namely the set of sites $\mathcal{S}$ and the corresponding observables y is associated with the root of a tree. A binary split divides $\mathcal{S}$ in two disjoint subsets $\mathcal{S}^{\rm left}$ and $\mathcal{S}^{\rm right}$, each with its subset of observables, associated with the root children. Recursion on the newly generated nodes, driven by a suitable parameter (*split gain*) to establish split priority/opportunity, produces a binary tree of classes and the associated segmentation of the image. At each node t, segmentation is carried out according to the MAP criterion

$$\widehat{x}^t = \arg\max_x p(x^t|y^t) = \arg\max_x p(y^t|x^t)p(x^t) \tag{1}$$

where $p(\cdot)$ indicates probability mass/density function (pmf or pdf) x^t is the label map, and $\widehat{x}^t$ is, therefore, the most probable map given the observables. The label map is modeled by a suitable MRF (see [18] for further details) to control the spatial regularity of $\widehat{x}^t$, while the data field y^t is assumed to be a spatially independent Gaussian given the labeling x^t, i.e., $p(y^t|x^t) = \prod_{s\in\mathcal{S}^t} p(y_s^t|x_s^t)$, with $y_s^t|x_s^t \sim N(\mu^t, \mathbf{C}^t)$.

R-TFR, an evolution of the TFR algorithm, is instead oriented to the segmentation of textured images. TFR, in particular, comprises three major processing steps:

1) spectral-based segmentation;
2) segment clustering;
3) progressive cluster merging.

The first step detects all elementary spectrally homogeneous-connected regions by means of any conventional region-based or edge-based segmenter. The second step forms clusters of segments that are similar in terms of spectral response, shape, and contextual interaction with the neighboring segments. The final step reconstructs the desired textures by progressive pair-wise merging of clusters. A suitable merging gain (called texture score) [20] is defined to decide which texture components should be merged at a given step. In particular, if we choose to stop this reconstruction process just before the last merging, we get a binary segmenter, namely the *binary*-TFR. Now, disposing of a texture-based binary segmentation engine, we can recursively apply it in a top-down fashion, following the same paradigm as TS-MRF, obtaining the recursive-TFR. The reader is referred to [19] for additional details on R-TFR.

DHC shares the split-wise paradigm with TS-MRF and R-TFR, and, in particular, it inherits their core binary segmentation engines: a binary-MRF and a binary-TFR. Each node/class is split in both ways and, according to a given criterion, the "best" split is accepted. Therefore, the overall segmentation process switches dynamically between two competing segmentation modeling types, a spectral-based and a texture-based one.

The DHC process is driven by proper "split gains," assigned locally to each node of the segmentation tree, which indicate both the priority of the node split and the most appropriate segmentation engine. The concept of split gain was introduced to control the TS-MRF evolution [18], and is basically a likelihood ratio between the split and the nonsplit hypotheses for the given node: given a partition of the data attached to the node of interest, the split gain balances the statistical fitting gain provided by the split with its representation cost. In particular, DHC defines two different gains by assessing the fitting gain with two different data models. The former $G^{S,t}$, where t indicates the associated node, is the original split gain used in TS-MRF, where each class is assumed to be normally distributed. In DHC, $G^{S,t}$ is used to assess the goodness of the spectral-based split (binary-MRF). $G^{T,t}$ is used, on the other hand, to score the texture-based split (binary-TFR), and assumes each class to be a Gaussian mixture (obviously, the hypothesis better suited for textures). For each node t, the model with the largest score is chosen and a score $G^t = \max\{G^{S,t}, G^{T,t}\}$ is associated with node t, fixing the split priority. Eventually, the segmentation proceeds, first splitting the leaves with the higher score. Accordingly, the output will be a sequence $S_2, S_3, \ldots, S_M$ of segmentations in $2, 3, \ldots, M$ classes, respectively, i.e., a hierarchical segmentation.

DHC, as well as TS-MRF and R-TFR, has been tested on the ALI benchmark in two configurations: 1) DHC/M, where the segmentation S_k is hand-picked from the hierarchical stack based on the number of regions singled out; and 2) DHC/K, where the number of classes k is given *a priori*.

B. Markovian Parameter Space Segmenter Family

MW3AR is an unsupervised multispectral, multiresolution, multiple-segmenter [22] for textured images with an unknown number of classes. The segmenter is based on a weighted combination of several unsupervised segmentation results, each in different resolution, using the modified sum rule. Multispectral textured image mosaics are locally represented by four causal directional multispectral random field models recursively evaluated for each pixel. A single local texture model is expressed as a stationary, causal, uncorrelated, noise-driven, 3-D, and autoregressive process [23]

$$Y_r = \gamma X_r + e_r \tag{2}$$

where $\gamma = [A_1, \ldots, A_\eta]$ is the parameter matrix, $r = [r_1, r_2]$ is the regular lattice multiindex, I_r^c is a causal neighborhood index set with $\eta = \; card(I_r^c)$, and e_r is a white Gaussian noise vector with zero mean and a constant but unknown covariance, X_r is the corresponding vector of the contextual neighbors Y_{r-s}. The single-resolution segmentation part of the algorithm is based on the underlying Gaussian mixture model and starts with an over-segmented initial estimation, which is adaptively modified until the optimal number of homogeneous texture segments is reached.

AR3D+EM method [24] is the simplified single-resolution, single-segmenter version of the MW3AR unsupervised segmenter.

C. Commercial Segmenters

eCognition software distributed by Trimble [5] provides a multiscale segmentation algorithm based on *region growing* whose detailed description is given in [25]. It starts with each pixel forming one image object or region. At each step, a pair of image objects is merged into one larger object. The merging decision is based on the similarity of adjacent image objects. A merging cost is hence defined by weighting this similarity measure with suitable shape priors (size, compactness, etc.). These costs represent a *degree of fitting*. At a given step of the procedure, the degree of fitting is evaluated for each couple of adjacent objects, and the fusion corresponding to the lower merging cost is performed if it is smaller than a given *least degree of fitting*. The procedure stops when no further merges are possible. Evidently, a smaller *least degree of fitting* allows fewer merges than a larger one. Therefore, the size of the resulting image objects will grow with the *least degree of fitting* value, which is why this parameter is often referred to as a scale parameter. This parameter, together with the weights related to the shape priors, allows the user to set the scale of interest. For evaluation, we refer to this technique as eCognition/M, to specify that segmentation at each step is achieved by manually choosing the most appropriate scale parameter, while all other parameters are fixed. This conceptually corresponds to a manual pick from a set of hierarchical segmentation maps.

ENVI (Environment for Visualizing Images) is a commercial software platform, distributed by Exelis [6], which is widely employed in RS applications. Among the several functionalities embedded, it also provides a region-based segmentation tool resorting to the proprietary algorithm described in [26]. In particular, ENVI provides a watershed-based segmentation algorithm where the topographic surface utilized is a modified gradient magnitude. An initial gradient computed on a properly (edge-preserving) filtered version of the input image is somehow made "uniform in scale" in scale through the density function of the gradient map. Hence, the user interactively chooses the optimal threshold for the modified gradient to which the watershed transform is eventually applied. Note that, unlike eCognition, the selected threshold only indirectly fixes the scale of the retrieved segmentation map, since no explicit control of the shape and size of objects is possible. As for eCognition, the use of different thresholds results in a set of nested segmentations. Therefore, this technique is referred to as ENVI/M in experiments, since the thresholds are adaptively chosen for each test image.

D. Supervised Methods

The compared supervised classifiers implemented in RapidMiner 5 [27] (kNN, UPGMA+kNN, 1NN, AM+kNN, Neuralnet) are, respectively, four variants of the k-nearest neighbor method and a multilayer feed-forward neural network segmenter, using local means or features selected from the hierarchical agglomerating unweighted pair group clustering (UPGMA), followed by the majority filter postprocessing.

IV. Comparative Analysis

To prove the effectiveness and large-scale potential of the proposed RS segmentation benchmark, we perform here a comparative analysis of several techniques by means of a critical reading of the related score tables available on the benchmark website. In particular, several unsupervised techniques from the

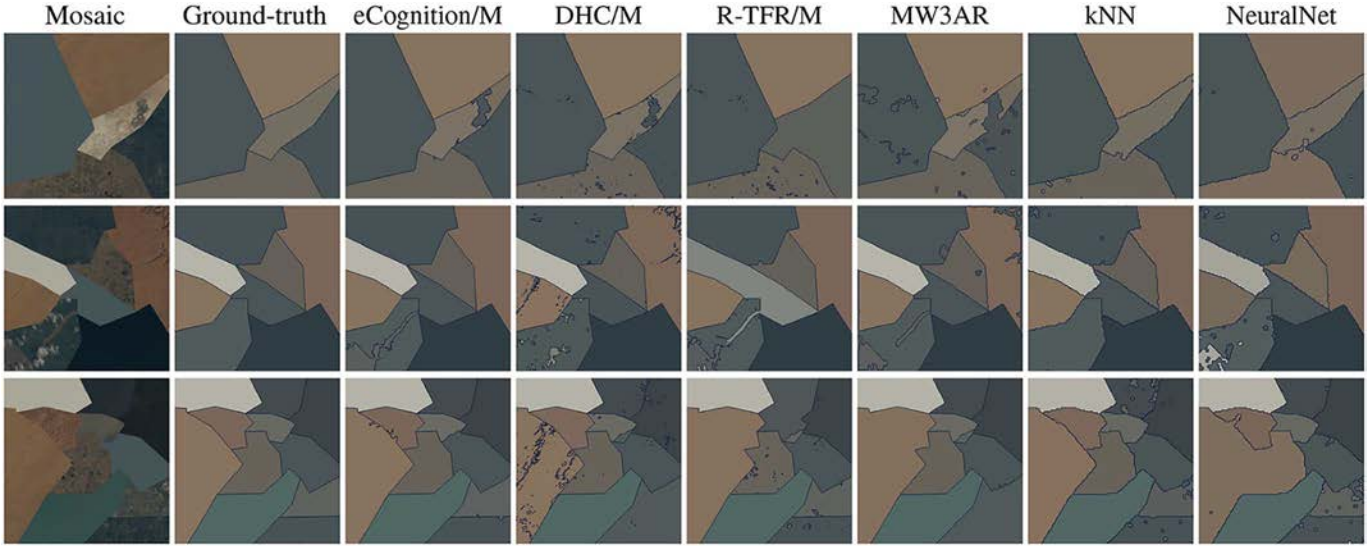


Fig. 1. Sample segmentation results for the ALI dataset. From left to right: the mosaic, the ground-truth, and the segmentations provided by eCognition/M, DHC/M, R-TFR/M, MW3AR, kNN, and NeuralNet.

two families defined in Section III are compared in detail, which differ significantly both in the approach and the methodology.

A. ALI Dataset

Fig. 1 shows segmentation results for three selected 512×512 mosaics from the ALI benchmark comprising from 5 to 11 multispectral satellite textures. The first two columns show the mosaics and their corresponding ground-truths. The remaining six show the segmentation maps provided by six alternative algorithms: 1) eCognition/M; 2) DHC/M; 3) R-TFR/M; 4) MW3AR; 5) kNN; and 6) NeuralNet. Note that the last two of these segmenters are supervised (kNN and NeuralNet).

Integrated numerical results over the whole normal ALI benchmark (ten different mosaics) are shown in Table II, where ↑ / ↓ denotes the required criterion direction, and bold numbers mark the best criterion value achieved from the 15 compared methods.

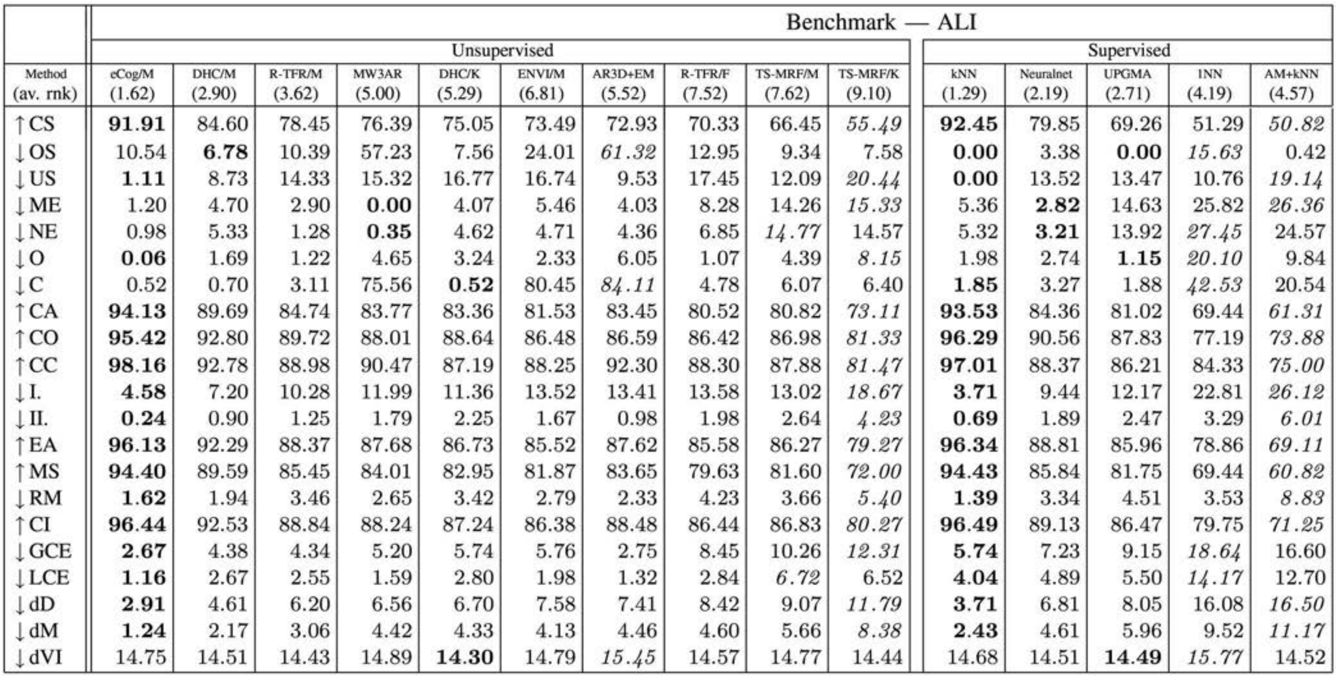
The results obtained using the eCognition commercial software achieve the highest ranking for most of the proposed figures. The quality of the corresponding segmentation maps is easily confirmed by visual inspection (see Fig. 1). However, despite the large performance gap with the other methods, it must be considered that eCognition is a region growing method, hence naturally providing connected component segmentations. Moreover, the fixed-scale approach allows for the extraction of regularly shaped objects that do not differ significantly in size. Both of these facts, which may not be appropriate in real-world cases, turn out to be advantageous here, where compact ground-truths generated through Voronoi tessellation are considered. As a matter of fact, eCognition provides a segmentation tool, which is particularly suitable for the proposed test images. For these reasons, we might reasonably consider eCognition as a performance “trendsetter” for unsupervised segmentation in this context. Nevertheless, it must also be pointed out that user interaction is made considerably heavy by the manual assessment of the shape weights (fixed for all test images) and the selection of the scale parameters (adaptively for each test image). This analysis is confirmed by the fact that results achieved using ENVI/M, which does not make use of any “direct” scale prior while adopting a similar connected component approach, are less satisfying. Over-segmentation is, in particular, more significant w.r.t. the top ranked techniques.

On the basis of the above-mentioned considerations, a more equitable comparison can be made among the other techniques. Standing on the overall scores of Table II, DHC/M ranks just below eCognition/M on most of the benchmark criteria and outperforms many of the supervised methods. In conformity with previous techniques of the same family of methods, which were previously also tested on the color texture benchmark [28], this technique once again proves to be effective in extracting large-scale textures. The choice of relying selectively on spectral and textural properties is particularly rewarding on the ALI dataset, where textured patches are often intertwined with areas in which the sole spectral information is more relevant. In numbers, DHC/M outperforms other techniques on region-based figures (CS/OS/US and class/object accuracies), helped by the selection of a segmentation map from the hierarchical stack whose scale matches that of the ground-truth. The lowest ranking criteria for DHC/M are error measures (ME/NE and LCE/GCE). This is mainly caused by the absence of a connected component approach, which leads to misclassification of smaller spectral/textural outliers. This clearly highlights the tradeoff between the accuracy at higher scales and the preservation of finer details.

The comparative analysis of DHC limited to its ancestors, R-TFR and TS-MRF, highlights one of the claimed objectives of the benchmark (“to check the progress of algorithm development”). In this case, the benchmark provided a quantitative answer to a fundamental question related to the DHC method: is it right to compare the two gains, $G^{S,t}$ and $G^{T,t}$, to decide whether to use a MRF or a TFR engine to split any node t? Regardless of the theoretical formulation, numerical evidence states that it is a good choice for the ALI dataset.

Both the techniques of the autoregressive model-based family, namely AR3D+EM and MW3AR, provide very

TABLE II
ALI BENCHMARK RESULTS FOR ECOG/M, DHC/M, R-TFR/M, MW3AR, DHC/K, ENVI/M, AR3D+EM, R-TFR/F, TS-MRF/M, TS-MRF/K, KNN, NEURALNET, UPGMA+KNN, 1NN, AND AM+KNN



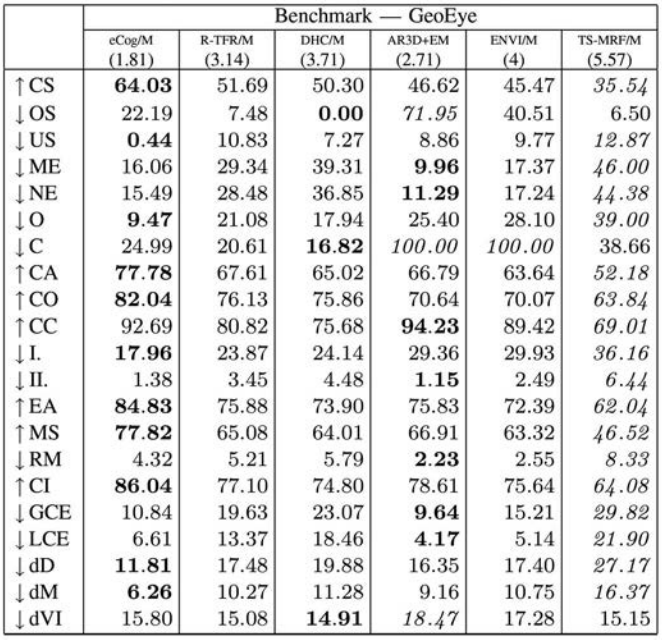
	Benchmark — ALI														
	Unsupervised										Supervised				
Method (av. rnk)	eCog/M (1.62)	DHC/M (2.90)	R-TFR/M (3.62)	MW3AR (5.00)	DHC/K (5.29)	ENVI/M (6.81)	AR3D+EM (5.52)	R-TFR/F (7.52)	TS-MRF/M (7.62)	TS-MRF/K (9.10)	kNN (1.29)	Neuralnet (2.19)	UPGMA (2.71)	1NN (4.19)	AM+kNN (4.57)
↑CS	**91.91**	84.60	78.45	76.39	75.05	73.49	72.93	70.33	66.45	*55.49*	**92.45**	79.85	69.26	51.29	*50.82*
↓OS	10.54	**6.78**	10.39	57.23	7.56	24.01	*61.32*	12.95	9.34	7.58	**0.00**	3.38	**0.00**	*15.63*	0.42
↓US	**1.11**	8.73	14.33	15.32	16.77	16.74	9.53	17.45	12.09	*20.44*	**0.00**	13.52	13.47	10.76	*19.14*
↓ME	1.20	4.70	2.90	**0.00**	4.07	5.46	4.03	8.28	14.26	*15.33*	5.36	**2.82**	14.63	25.82	*26.36*
↓NE	0.98	5.33	1.28	**0.35**	4.62	4.71	4.36	6.85	*14.77*	14.57	5.32	**3.21**	13.92	*27.45*	24.57
↓O	**0.06**	1.69	1.22	4.65	3.24	2.33	6.05	1.07	4.39	*8.15*	1.98	2.74	**1.15**	*20.10*	9.84
↓C	0.52	0.70	3.11	75.56	**0.52**	80.45	*84.11*	4.78	6.07	6.40	**1.85**	3.27	1.88	*42.53*	20.54
↑CA	**94.13**	89.69	84.74	83.77	83.36	81.53	83.45	80.52	80.82	*73.11*	**93.53**	84.36	81.02	69.44	*61.31*
↑CO	**95.42**	92.80	89.72	88.01	88.64	86.48	86.59	86.42	86.98	*81.33*	**96.29**	90.56	87.83	77.19	*73.88*
↑CC	**98.16**	92.78	88.98	90.47	87.19	88.25	92.30	88.30	87.88	*81.47*	**97.01**	88.37	86.21	84.33	*75.00*
↓I.	**4.58**	7.20	10.28	11.99	11.36	13.52	13.41	13.58	13.02	*18.67*	**3.71**	9.44	12.17	22.81	*26.12*
↓II.	**0.24**	0.90	1.25	1.79	2.25	1.67	0.98	1.98	2.64	*4.23*	**0.69**	1.89	2.47	3.29	*6.01*
↑EA	**96.13**	92.29	88.37	87.68	86.73	85.52	87.62	85.58	86.27	*79.27*	**96.34**	88.81	85.96	78.86	*69.11*
↑MS	**94.40**	89.59	85.45	84.01	82.95	81.87	83.65	79.63	81.60	*72.00*	**94.43**	85.84	81.75	69.44	*60.82*
↓RM	**1.62**	1.94	3.46	2.65	3.42	2.79	2.33	4.23	3.66	*5.40*	**1.39**	3.34	4.51	3.53	*8.83*
↑CI	**96.44**	92.53	88.84	88.24	87.24	86.38	88.48	86.44	86.83	*80.27*	**96.49**	89.13	86.47	79.75	*71.25*
↓GCE	**2.67**	4.38	4.34	5.20	5.74	5.76	2.75	8.45	10.26	*12.31*	**5.74**	7.23	9.15	*18.64*	16.60
↓LCE	**1.16**	2.67	2.55	1.59	2.80	1.98	1.32	2.84	*6.72*	6.52	**4.04**	4.89	5.50	*14.17*	12.70
↓dD	**2.91**	4.61	6.20	6.56	6.70	7.58	7.41	8.42	9.07	*11.79*	**3.71**	6.81	8.05	16.08	*16.50*
↓dM	**1.24**	2.17	3.06	4.42	4.33	4.13	4.46	4.60	5.66	*8.38*	**2.43**	4.61	5.96	9.52	*11.17*
↓dVI	14.75	14.51	14.43	14.89	**14.30**	14.79	*15.45*	14.57	14.77	14.44	14.68	14.51	**14.49**	*15.77*	14.52

Benchmark criteria: CS, correct segmentation; OS, over-segmentation; US, under-segmentation; ME, missed error; NE, noise error; O, omission error; C, commission error; CA, class accuracy; CO, recall—correct assignment; CC, precision—object accuracy; I, type I error; II, type II error; EA, mean class accuracy estimate; MS, mapping score; RM, root mean square proportion estimation error; CI, comparison index; GCE, global consistency error; LCE, local consistency error; dD, Van Dongen metric; dM, Mirkin metric; dVI, variation of information.

interesting error measures, which are worth a deeper insight. Despite the high over-segmentation (OS criterion), AR3D+EM ranks first on both the LCE and GCE indicators. These consistency figures are known to provide indications on how much a given segmentation map can be considered the refinement of the reference ground-truth. However, in the general case, a deeply over-segmented map can achieve very low LCE and GCE values; hence, we need to consider other figures to validate this possible qualifying point. A positive counter-check is first given by region based error measures (ME/NE), on which AR3D+EM exhibits a very good score, confirming that no significant misclassification has taken place. The good value of the precision measure (CC) is another confirmation that most of the extracted details are relevant, although the low recall (CO) indicates that some others are missing.

The MW3AR technique shows a significant improvement of the performance compared to its predecessor, providing outstanding values for the region-based error measures (ME/NE), which indicate a distinct potential in avoiding wrong contours. These two indicators are somehow complementary to the other region-based figures (CS/OS/US) [14]; it can be interesting to observe how values for this class of indicators have been redistributed with respect to AR3D+EM. It is immediate to notice that the higher CS is accompanied by a lower OS and more significant under-segmentation (US). This behavior is typical of segmentation particularly sensitive to the scale of objects: in principle, if the scale of segmentation is fixed, objects above or below this targeted scale are, respectively, under- and over-segmented. This observation is confirmed by the visual inspection of results, whereas the results provided by MW3AR exhibit a more uniform scale compared to AR3D+EM, mainly in terms of spectral differences.

TABLE III
GEOEYE BENCHMARK RESULTS FOR ECOG/M, R-TFR/M, DHC/M, AR3D+EM, ENVI/M, TS-MRF/M



	Benchmark — GeoEye					
	eCog/M (1.81)	R-TFR/M (3.14)	DHC/M (3.71)	AR3D+EM (2.71)	ENVI/M (4)	TS-MRF/M (5.57)
↑CS	**64.03**	51.69	50.30	46.62	45.47	*35.54*
↓OS	22.19	7.48	**0.00**	*71.95*	40.51	6.50
↓US	**0.44**	10.83	7.27	8.86	9.77	*12.87*
↓ME	16.06	29.34	39.31	**9.96**	17.37	*46.00*
↓NE	15.49	28.48	36.85	**11.29**	17.24	*44.38*
↓O	**9.47**	21.08	17.94	25.40	28.10	*39.00*
↓C	24.99	20.61	**16.82**	*100.00*	*100.00*	38.66
↑CA	**77.78**	67.61	65.02	66.79	63.64	*52.18*
↑CO	**82.04**	76.13	75.86	70.64	70.07	*63.84*
↑CC	92.69	80.82	75.68	**94.23**	89.42	*69.01*
↓I.	**17.96**	23.87	24.14	29.36	29.93	*36.16*
↓II.	1.38	3.45	4.48	**1.15**	2.49	*6.44*
↑EA	**84.83**	75.88	73.90	75.83	72.39	*62.04*
↑MS	**77.82**	65.08	64.01	66.91	63.32	*46.52*
↓RM	4.32	5.21	5.79	**2.23**	2.55	*8.33*
↑CI	**86.04**	77.10	74.80	78.61	75.64	*64.08*
↓GCE	10.84	19.63	23.07	**9.64**	15.21	*29.82*
↓LCE	6.61	13.37	18.46	**4.17**	5.14	*21.90*
↓dD	**11.81**	17.48	19.88	16.35	17.40	*27.17*
↓dM	**6.26**	10.27	11.28	9.16	10.75	*16.37*
↓dVI	15.80	15.08	**14.91**	*18.47*	17.28	15.15

For benchmark criteria, see Table II.

Finally, recall that neither of these methods makes use of any provided information, which further attests to the quality of the

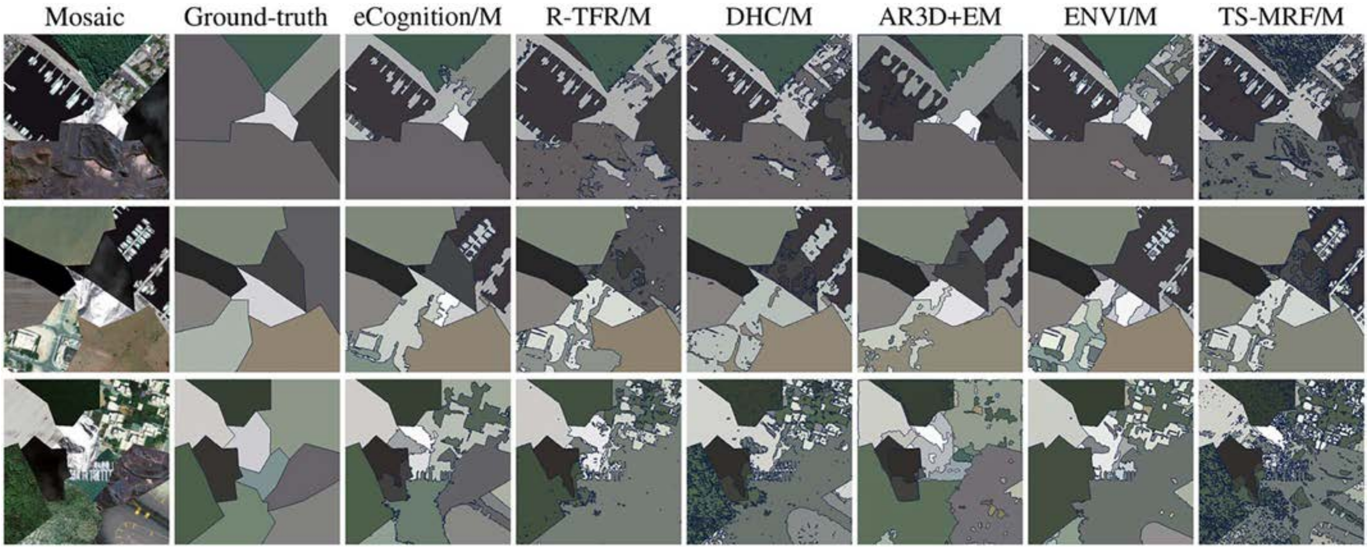


Fig. 2. Sample segmentation results for the GeoEye dataset. From left to right: the mosaic, the ground-truth, and the segmentations provided by eCog/M, R-TFR/M, DHC/M, AR3D+EM, ENVI/M, and TS-MRF/M.

results shown here. In conclusion, AR3D+EM shows a good potential for finer segmentations, which could either be applied to object layer extraction or integrated into a hierarchical framework. MW3AR successfully addresses the problems of over-segmentation of AR3D+EM, achieving a better overall score at the price of a reduced flexibility in detecting contours among regions at different scales.

A similar analysis can be conducted to make comparisons with other techniques currently reported on the benchmark websites, both supervised and unsupervised, which will be left to the reader to explore.

A final remark concerns supervised segmentation techniques: their results are, unexpectedly, globally less accurate than for unsupervised ones, except for the technique denoted as kNN. This outcome is due to the fact that they all use oversimplified features (local arithmetic averages), which cannot compete with the state-of-the-art textural features (color Markovian, LBP) utilized by the unsupervised segmenters. This result demonstrates the importance of textural representation for modern high-resolution RS data. Similarly, for the top scoring kNN technique, the decision rule for clustering is less precise than the Gaussian-mixture-model-based used in MW3AR and AR3D+EM methods, as testified by the higher error values (ME/NE/LCE/GCE) reached using these techniques. In this case, the benchmark also shows its potential in highlighting room left for future research, which is probably the ultimate finality for this class of tools.

B. GeoEye Dataset

The GeoEye dataset has been recently introduced to extend the capabilities of the benchmark to sensors providing very high spatial resolution images. This dataset is currently experimental, and is expected to evolve in the near future.

Switching the resolution up to one-tenth of a meter, one must deal with the occurrence of very long range textural patterns and out-of-scale objects, which makes the extraction of regions of interest a particularly challenging task. In result, all the benchmarked techniques perform uniformly worse than on the ALI dataset, as shown in Table III. Segmentation maps provided by several sample methods are depicted in Fig. 2.

Again, the eCognition segmentation tool performs best for this type of image. This further proves that, in the absence of a proper high-level modeling of complex scenes, the prior density ensuring the desired scale/compactness features of this method better matches the characteristics of the test images. Needless to say, the assessment of parameters has required a greater effort w.r.t. other techniques.

Evidently, spectral-based techniques such as TSMRF/M perform poorly on these datasets. Methods which introduce a fine-to-coarse texture modeling, such as ENVI/M and especially AR3D+EM, can progressively better capture interactions among image elements, providing a better CS and low error measures (ME/NE/LCE/GCE). A higher region-based accuracy can be achieved with DHC/M and R-TFR/M, which specifically account for long-range textures, at the price of higher error figures. R-TFR/M, in particular, outperforms DHC/M, which in the absence of robust textural information tends to rely on a spectral-based segmentation, which is unsuitable for these images.

V. Conclusion

The implemented supervised/unsupervised RS segmentation benchmark is the fully automatic Web application, which enables us, for the first time, to objectively compare image segmentation algorithms on extensive test sets, thereby providing an important tool for the progress of new segmentation methods. RS classifiers can be ranked based on a best-fitting criterion chosen from the set of 27 distinct criteria. Test mosaics as well as ground-truths are automatically generated, which both guarantees the objective evaluation and the easy generation of extensive test sets which are otherwise infeasible to achieve. PTSDB verifies single algorithms against others on multispectral or RGB ALI and high-resolution color GeoEye satellite data and tests their noise resistance. The researchers can quickly and effectively compare their progress and check their performance characteristics.

Further developments are currently being carried out to address several issues. On one hand, the generation of ground-truths, which better approximate real circumstances (varying scale, spatial distribution, and shape of regions of interest), is being addressed. Moreover, for the newly introduced GeoEye dataset, the generation of bigger images, to favor the emergence of texture patterns, is being taken into account.

ACKNOWLEDGMENT

The authors would like to thank the Faculty of Information Technology, CTU students J. Krejcar, V. Medonos, M. Ovesný, M. Dvorožňák, and G. Masi for providing experiments with the kNN, NeuralNet, UPGMA+kNN, 1NN, AM+kNN, ENVI, and eCognition algorithms.

REFERENCES

[1] P. Li, J. Guo, B. Song, and X. Xiao, "A multilevel hierarchical image segmentation method for urban impervious surface mapping using very high resolution imagery," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 4, no. 1, pp. 103–116, Mar. 2011.

[2] T. R. Martha, N. Kerle, C. J. van Westen, V. Jetten, and K. V. Kumar, "Segment optimization and data-driven thresholding for knowledge-based landslide detection by object-based image analysis," *IEEE Trans. Geosci. Remote Sens.*, vol. 49, no. 12, pp. 4928–4943, Dec. 2011.

[3] B. W. Heumann, "An object-based classification of mangroves using a hybrid decision tree—Support vector machine approach," *Remote Sens.*, vol. 3, no. 12, pp. 2440–2460, Nov. 2011.

[4] T. Blaschke, "Object based image analysis for remote sensing," *ISPRS J. Photogramm. Remote Sens.*, vol. 65, no. 1, pp. 2–16, 2010.

[5] Trimble. (2012). *eCognition Software* [Online]. Available: http://www.ecognition.com

[6] Exelis. (2012). *ENVI* [Online]. Available: http://www.exelisvis.com

[7] J. Yuan, D. Wang, and R. Li, "Remote sensing image segmentation by combining spectral and texture features," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 1, pp. 16–24, Jan. 2014.

[8] C. Kurtz, N. Passat, P. Gançarski, and A. Puissant, "Extraction of complex patterns from multiresolution remote sensing images: A hierarchical top-down methodology," *Pattern Recognit.*, vol. 45, no. 2, pp. 685–706, 2012.

[9] L. Yi, G. Zhang, and Z. Wu, "A scale-synthesis method for high spatial resolution remote sensing image segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 50, no. 10, pp. 4062–4070, Oct. 2012.

[10] J. A. dos Santos, P.-H. Gosselin, S. Philipp-Foliguet, R. S. Torres, and A. X. Falcão, "Multiscale classification of remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 50, no. 10, pp. 3764–3775, Oct. 2012.

[11] I. Epifanio and P. Soille, "Morphological texture features for unsupervised and supervised segmentations of natural landscapes," *IEEE Trans. Geosci. Remote Sens.*, vol. 45, no. 4, pp. 1074–1083, Apr. 2007.

[12] M. Haindl and S. Mikeš, "Texture segmentation benchmark," in *Proc. 19th Int. Conf. Pattern Recogn. (ICPR'08)*, Dec. 2008, pp. 1–4 [Online]. Available: http://doi.ieeecomputersociety.org/

[13] GeoEye, Inc. (2009). *Geoeye Product Guide v1.0.1* [Online]. Available: http://www.genesiis.com/pdf/GeoEye-1-product-guide.pdf

[14] A. Hoover *et al.*, "An experimental comparison of range image segmentation algorithms," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 18, no. 7, pp. 673–689, Jul. 1996.

[15] D. Martin, C. Fowlkes, D. Tal, and J. Malik, "A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics," in *Proc. 8th Int. Conf. Comput. Vis.*, vol. 2, Jul. 2001, pp. 416–423 [Online]. Available: http://www.cs.berkeley.edu/projects/vision/grouping/segbench/

[16] M. Meila, "Comparing clusterings—An axiomatic view," in *Proc. 7th Int. Conf. Mach. Learn. (ICML)*, 2005, pp. 577–584.

[17] G. Scarpa, G. Masi, R. Gaetano, L. Verdoliva, and G. Poggi, "Dynamic hierarchical segmentation of remote sensing images," in *Image Analysis and Processing*, vol. 8156, A. Petrosino, Ed. New York, NY, USA: Springer, 2013, pp. 371–380.

[18] C. D'Elia, G. Poggi, and G. Scarpa, "A tree-structured Markov random field model for Bayesian image segmentation," *IEEE Trans. Image Process.*, vol. 12, no. 10, pp. 1259–1273, Oct. 2003.

[19] R. Gaetano, G. Scarpa, and G. Poggi, "Recursive texture fragmentation and reconstruction segmentation algorithm applied to VHR images," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS'09)*, vol. 4, 2009, pp. IV–101–IV–104.

[20] G. Scarpa and M. Haindl, "Unsupervised texture segmentation by spectral-spatial-independent clustering," in *Proc. Int. Conf. Pattern Recogn.*, 2006, pp. 151–154.

[21] S. Z. Li, *Markov Random Field Modeling in Image Analysis*, 3rd ed. New York, NY, USA: Springer, 2009.

[22] M. Haindl, S. Mikeš, and P. Pudil, "Unsupervised hierarchical weighted multi-segmenter," in *Lecture Notes in Computer Science*, vol. 5519, J. Benediktsson, J. Kittler, and F. Roli, Eds. New York, NY, USA: Springer, 2009, pp. 272–282.

[23] M. Haindl, "Visual data recognition and modeling based on local Markovian models," in *Mathematical Methods for Signal and Image Analysis and Representation*, vol. 41, L. Florack, R. Duits, G. Jongbloed, M.-C. Lieshout, and L. Davies, Eds. New York, NY, USA: Springer, 2012, ch. 14, pp. 241–259.

[24] M. Haindl, S. Mikeš, and P. Vácha, "Illumination invariant unsupervised segmenter," in *Proc. IEEE 16th Int. Conf. Image Process. (ICIP'09)*, 2009, pp. 4025–4028.

[25] M. Baatz and A. Schäpe, "Multiresolution Segmentation: An optimization approach for high quality multi-scale image segmentation," in *Proc. Angew. Geogr. Inf. XII. Beiträge zum AGIT-Symp.*, 2000, pp. 12–23.

[26] J. Xiaoying, "Segmentation-based image processing system," U.S. Patent App. 11/984,222, May 14, 2009 [Online]. Available: http://www.google.com/patents/US20090123070

[27] F. Akthar and C. Hahne, "Rapidminer 5 operator reference," Rapid-I GmbH, 2012.

[28] G. Scarpa, M. Haindl, and J. Zerubia, "A hierarchical finite-state model for texture segmentation," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2007, pp. I–1209–I–1212.

Stanislav Mikeš received the M.Sc. (Mgr.) degree in computer science and the Ph.D. degree in unsupervised image segmentation from the Faculty of Mathematics and Physics, Charles University, Prague in 2002 and 2010, respectively.

Since 2002, he has been a Researcher with the Department of Pattern Recognition, Institute of Information Theory and Automation, Czech Academy of Sciences, Prague, Czech Republic. His research interests include unsupervised image segmentation, texture classification, verification methodology and benchmarking, and virtual reality.

Michal Haindl (SM'95) received the Degree in control engineering from the Czech Technical University, Prague, Czech Republic, in 1979, the Ph.D. degree in technical cybernetics from the Czechoslovak Academy of Sciences, Prague, Czech Republic, in 1983, and the Sc.D. (Dr.Sc.) degree from the Czech Technical University in 2001.

He is a Professor with International Association for Pattern Recognition (IAPR), Durham, NC, USA. From 1983 to 1990, he worked with the Institute of Information Theory and Automation, Czechoslovak Academy of Sciences, Prague, Czech Republic, on different adaptive control, image processing, and pattern recognition problems. From 1990 to 1995, he was with the University of Newcastle, Callaghan N.S.W., Australia; Rutherford Appleton Laboratory, Didcot, U.K.; Centre for Mathematics and Computer Science, Amsterdam, The Netherlands, and Institute National de Recherche en Informatique et en Automatique, Rocquencourt, France, working on several image analysis and pattern recognition projects. In 1995, he rejoined the Institute of Information Theory and Automation, where he is the Head of the Pattern Recognition Department. He is the author of about 300 research papers published in books, journals, and conference proceedings. His research interests include random fields applications in pattern recognition and image processing and automatic acquisition of virtual reality models.

Dr. Haindl is a fellow of the IAPR. He is an Associate Editor of the *International Journal of Pattern Recognition and Artificial Intelligence*, *Kybernetika*, and has served on the program committees of numerous conferences.

Giuseppe Scarpa (M’12) received the Laurea (M.S.) degree in telecommunication engineering and the Ph.D. degree in electronic and telecommunication engineering from the University Federico II, Naples, Italy, in 2001 and 2005, respectively.

He was a Visiting Student at INRIA, France. Thanks to a joint ERCIM Postdoc Fellowship, he was awarded in 2004, he has been a Research Fellow with both the UTIA, Czech Academy of Sciences, Prague, Czech Republic, in 2005, and the INRIA Institute, in 2006. Since 2006, he has been an Assistant Professor with the Department of Electrical Engineering and Information Technology, University Federico II. His research interests include image analysis, and, in particular, segmentation, texture modeling and classification, object detection, and filtering, with applications in both remote sensing and medical domains.

Prof. Scarpa was the recipient of the Marie Curie Scholarship Award in 2003. He is currently an Associate Editor of the IEEE SIGNAL PROCESSING LETTERS.

Raffaele Gaetano received the Laurea (M.S.) degree in computer engineering and the Ph.D. degree in electronic and telecommunication engineering from the University Federico II, Naples, Italy, in 2004 and 2009, respectively.

He has been a ERCIM Post-Doctoral Fellow of both the ARIANA team of INRIA Sophia Antipolis and the DEVA team of SZTAKI, Hungarian Academy of Sciences. From 2010 to 2013 he has been a Postdoctoral Fellow with TELECOM PARISTECH, France, within the MultiMédia Group. He is currently a Postdoc Member of the Research Group on Image Processing (GRIP), Department of Electrical Engineering and Information Technology, University Federico II. His research interests include the field of image and video analysis and processing, as well as color- and texture-based hierarchical image segmentation, morphological image analysis, and object detection, mainly applied to the classification of remote sensing images, to image restoration, stereo vision, and image/video super-resolution.