Remote Sensing Segmentation Benchmark

Stanislav Mikeš Michal Haindl
Institute of Information Theory and Automation
of the ASCR, 182 08 Prague, Czech Republic
{*xaos,haindl*}*@utia.cz*

Giuseppe Scarpa
University Federico II of Naples
Via Claudio 21, 80125, Naples, Italy
giscarpa@unina.it

Abstract

*In this work we present the enrichment of the Prague texture segmentation data-generator and benchmark (PTSDB) also for the assessment of the remote sensing 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 remote sensing datasets contain ten-spectral ALI satellite images and their RGB subsets, with optional additive noise resistance checking. Alternative setting options allow to test also scale, rotation or illumination invariance. The benchmark functionality is demonstrated by testing and comparing six remote sensing segmentation algorithms.*

1. Introduction

Satellite image segmentation is the prerequisite for successful remote sensing scene analysis, used, for example, in crop inventory, geological and environment surveys, military applications, etc. Although a large number of methods were already published [1, 2], this problem is still far from being solved. This is also due to the lack of reliable and objective means to compare the performance of different techniques. Very limited efforts were made, in fact, to develop suitable quantitative measures of segmentation quality, especially in the case of remote sensing. In this field, in fact, it is 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. Although this is partially justified by the large number of data sources available and the many different applications of segmentation, this habit 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 alternative 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. Since no benchmark oriented to the development of segmentation methods for remote sensing is available, we have implemented a solution in the form of web based data generator and benchmark software. Proper testing and robust learning of performance characteristics require large test sets and objective ground truths which is unrealistic for natural satellite images. Thus, inevitably few used satellite test images 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 approximate satellite scenes. The most appealing feature of this compromise is the unlimited number of different test images with corresponding objective and free ground truth map available for each of them.

2. Benchmark

The Prague texture segmentation data-generator and benchmark (PTSDB) is a web based (**http://mosaic.utia.cas.cz**) service [4] 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 benchmark is to compute several accuracy measures for each given algorithm over the selected dataset. Once collected different segmentations over a given dataset, it is therefore possible to score them with respect to any of the computed accuracy indicators. This is of critical importance for

three main reasons:

1. to check the progress of an algorithm development,
2. to compare any method to any other,
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 data set and employing several evaluation measures for 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 non-informative positive examples, our benchmark possesses all these mentioned important features. While the colour benchmark textures were chosen on purpose to produce unusually difficult tests in order to leave large margins for future better segmentation algorithms, the ALI multispectral textures contain richer spectral information and thus their textural analysis is less demanding. The benchmark operates either in full mode for registered users (unrestricted mode - U) or in a restricted mode. The benchmark allows: to obtain customized experimental satellite texture mosaics and their corresponding ground truth (U); to obtain the benchmark mosaic sets with their corresponding ground truth; to evaluate working segmenters and compare them with the state-of-the-art methods; to update the benchmark database (U) with an algorithm details; to assess noise robustness; to check single mosaics evaluation details (criteria values and resulting thematic maps); to rank segmentation algorithms according to the most common benchmark criteria; to obtain LaTeX or MATLAB coded result tables (U); to select user-defined subset of criteria (U).

2.1. Remote Sensing Data

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

The remote sensing benchmark uses the Advanced Land Imager (ALI) 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 (SCA) and also covers 3° by 1.625°. 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 $1.20-1.30\ \mu m$. 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-meter resolution for the multispectral pixels and 10-meter resolution for the panchromatic pixels. ALI bands and spectral ranges are listed in Tab.1.

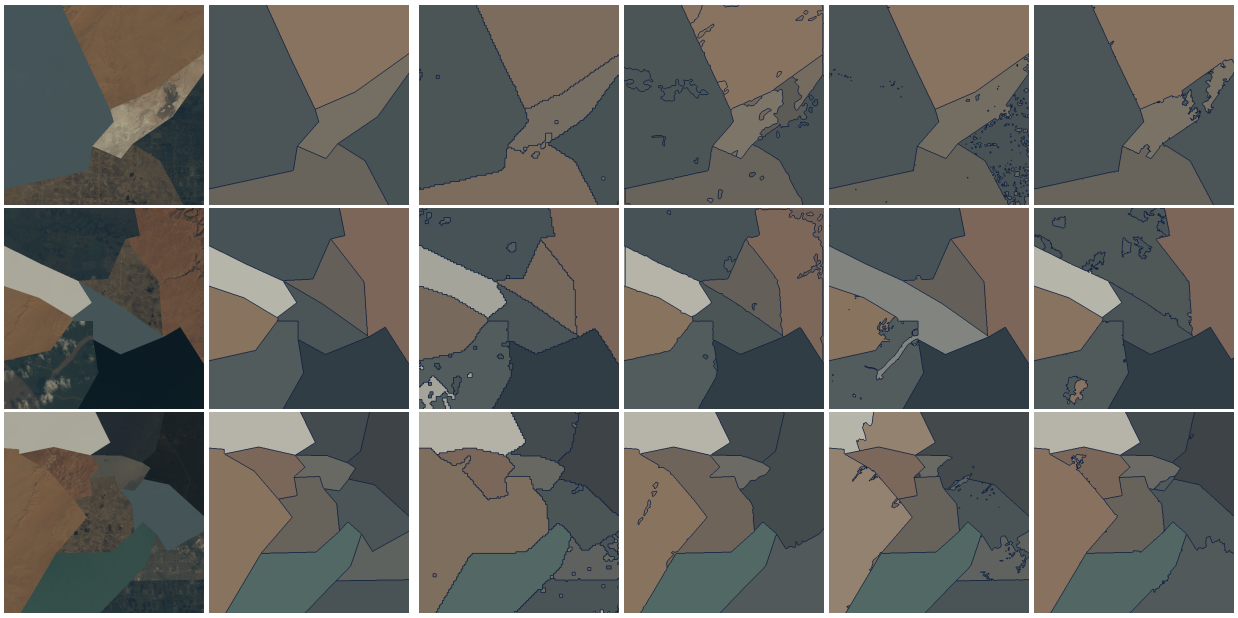
Table 1. ALI bands and spectral ranges.

	Band	Spectral Range $[\mu m]$	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

The benchmark uses 31 multispectral ALI textures categorized into 12 thematic classes. The satellite texture parts which are not used in the corresponding test mosaics are used as separate training sets in the benchmark-supervised mode.

2.2. Benchmark Sets Creation

Benchmark 512×512 test mosaics are built by means of a Voronoi polygon random generator, and filled with randomly selected ALI textures. It is worth emphasizing that smaller and irregularly shaped objects are more difficult to segment than usually used bigger and regular (squares or circles) objects. ALI benchmarks (multispectral and RGB) are generated upon request in three quantities (10, 40, 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 remote sensing benchmark allows to check the segmenter noise resistance. All generated mosaics can be corrupted with additive



a b c d e f

Figure 1. Selected benchmark texture mosaics (a), ground-truth (b), Neuralnet (c), AR3D+EM (d), R-TFR (e), UPGMA+kNN (f) segmentation results, respectively.

Gaussian, Poisson, or salt & pepper noise. Alternative benchmarks allow to test also scale, and rotation or illumination invariance of the evaluated segmentation algorithm.

3. 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 twenty seven evaluation criteria assorted into four thematic groups: region-based (5+5), pixel-wise (11+1), consistency measures (2) and clustering comparison criteria (3). The performance criteria mutually compare ground truth image regions with the corresponding machine segmented regions. The basic region-based criteria available 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, mapping score, etc. The consistency criteria are global and local consistency errors. Finally, the last set contains three clustering comparison measures.

4. Examples

The remote sensing ALI benchmark performance is demonstrated by comparing two unsupervised (our previously published methods AR3D+EM [5] and R-TFR [3]) and several supervised segmentation algorithms. The detailed performance of all these methods can be found on the benchmark server.

Fig. 1 shows segmentation results for three selected 512×512 mosaics from the ALI benchmark comprising from five to eleven multispectral satellite textures. The first two columns show the mosaics and their corresponding ground-truths. The remaining four columns show the segmentation maps provided by four alternative algorithms – Neuralnet, AR3D+EM, R-TFR, UPGMA+kNN. Two of these segmenters are supervised (Neuralnet, UPGMA+kNN) and the other two (AR3D+EM, R-TFR) are unsupervised.

Visual comparison suggests over-segmentation inclination of the AR3D+EM [5] algorithm which is confirmed by the objective evaluation criterion Tab.2. On the other hand this method outperforms all others in terms of correct localization of the region borders. Both unsupervised methods are comparable or even better than the supervised ones. The first four methods have at least one best performing criterion which suggests the optimal application for the corresponding methods. In-

Table 2. ALI benchmark results for Neuralnet, AR3D+EM, R-TFR, UPGMA+kNN, AM+DT, AM+kNN; (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).

	Benchmark – ALI					
	Neuralnet (2.11)	AR3D+EM (2.37)	R-TFR (3.22)	UPGMA+kNN (2.96)	AM+DT (4.93)	AM+kNN (5.37)
↑ *CS*	**79.85**	72.93	72.26	69.26	51.29	*50.82*
↓ *OS*	3.38	*61.32*	**0.00**	**0.00**	15.63	0.42
↓ *US*	13.52	**9.53**	*21.87*	13.47	10.76	19.14
↓ *ME*	**2.82**	4.03	4.35	14.63	25.82	*26.36*
↓ *NE*	3.21	4.36	**2.54**	13.92	*27.45*	24.57
↓ *O*	2.74	6.05	2.01	**1.15**	*20.10*	9.84
↓ *C*	3.27	*84.11*	5.04	**1.88**	42.53	20.54
↑ *CA*	**84.36**	83.45	77.70	81.02	69.44	*61.31*
↑ *CO*	**90.56**	86.59	85.62	87.83	77.19	*73.88*
↑ *CC*	88.37	**92.30**	79.81	86.21	84.33	*75.00*
↓ *I.*	**9.44**	13.41	14.38	12.17	22.81	*26.12*
↓ *II.*	1.89	**0.98**	2.21	2.47	3.29	*6.01*
↑ *EA*	**88.81**	87.62	81.48	85.96	78.86	*69.11*
↑ *MS*	**85.84**	83.65	78.43	81.75	69.44	*60.82*
↓ *RM*	3.34	**2.33**	5.04	4.51	3.53	*8.83*
↑ *CI*	**89.13**	88.48	82.07	86.47	79.75	*71.25*
↓ *GCE*	7.23	**2.75**	4.37	9.15	*18.64*	16.60
↓ *LCE*	4.89	**1.32**	3.28	5.50	*14.17*	12.70
↓ *dD*	**6.81**	7.41	8.55	8.05	16.08	*16.50*
↓ *dM*	4.61	**4.46**	4.68	5.96	9.52	*11.17*
↓ *dVI*	14.51	15.45	**13.92**	14.49	*15.77*	14.52

tegrated numerical results over the whole normal ALI benchmark (10 different mosaics) in Tab.2 (↑ / ↓ denotes the required criterion direction, bold numbers the best criterion value achieved) confirm these observations.

5. Conclusions

The implemented supervised / unsupervised remote sensing segmentation benchmark is the fully automatic web application which enables 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. Remote sensing classifiers can be ranked based on a chosen most fitting criterion from the set of twenty seven distinct criteria. The both test mosaics as well as the ground truths are computer created which guarantees not only the evaluation objectivity but simultaneously allows easy generation of extensive test sets which are otherwise infeasible to achieve.

PTSDB verifies single algorithms against others on multispectral or RGB ALI satellite data and tests their noise resistance. The researchers can quickly and effectively compare their progress and check their performance characteristics. Other applications such as machine learning, feature selection, image compression, QBIC methods evaluation, scale, and rotation or illumination invariance, etc., can advance from the PTSDB benchmark services as well.

Acknowledgments

This research was supported by grant GAČR 102/08/0593 and partially by the grants CESNET 409/2011, GAČR 103/11/0335. The authors wish to thank Faculty of information technology, CTU students J. Krejcar, V. Medonos, M. Ovesný, and M. Dvorožňák for providing experiments with the Neuralnet, UPGMA+kNN, AM+DT, AM+kNN supervised classifiers.

References

[1] P. Arbelàez, M. Maire, C. Fowlkes, and J. Malik. Contour detection and hierarchical image segmentation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 33(5):898 –916, may 2011.

[2] I. Epifanio and P. Soille. Morphological texture features for unsupervised and supervised segmentations of natural landscapes. *Geoscience and Remote Sensing, IEEE Transactions on*, 45(4):1074 –1083, april 2007.

[3] R. Gaetano, G. Scarpa, and G. Poggi. Recursive texture fragmentation and reconstruction segmentation algorithm applied to vhr images. In *Geoscience and Remote Sensing Symposium, 2009 IEEE International, IGARSS 2009*, volume 4, pages IV–101. IEEE, 2009.

[4] M. Haindl and S. Mikeš. Texture segmentation benchmark. In B. Lovell, D. Laurendeau, and R. Duin, editors, *Proceedings of the 19th International Conference on Pattern Recognition, ICPR 2008*, pages 1–4, Los Alamitos, December 2008. IEEE Computer Society.

[5] M. Haindl, S. Mikeš, and P. Vácha. Illumination invariant unsupervised segmenter. In M. Bayoumi, editor, *IEEE 16th Int. Conf. on Image Processing - ICIP 2009*, pages 4025–4028. IEEE, 2009.