Texture Segmentation Benchmark

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

Abstract

*The Prague texture segmentation data-generator and benchmark is a web based (***http://mosaic.utia.cas.cz***) service designed to mutually compare and rank different texture segmenters, and to support new segmentation and classification methods development. The benchmark verifies their performance characteristics on monospectral, multispectral, bidirectional texture function (BTF) data and enables to test their noise robustness, scale, and rotation or illumination invariance. It can easily be used for other applications such as feature selection, image compression, and query by pictorial example, etc. The benchmark functionalities are demonstrated on five previously published image segmentation algorithms evaluation.*

1. Introduction

Unsupervised or supervised texture segmentation is the prerequisite for successful content-based image retrieval, scene analysis, automatic acquisition of virtual models, quality control, security, medical applications and many others. Although more than 1000 different methods were already published [15], this problem is still far from being solved. This is among others due to missing reliable performance comparison between different techniques because very limited effort was spent to develop suitable quantitative measures of segmentation quality that can be used to evaluate and compare segmentation algorithms. Rather than advancing the most promising image segmentation approaches novel algorithms are often satisfied just being sufficiently different from the previously published ones and tested only on a few carefully selected positive examples. The optimal alternative to check several variants of a developed method and to carefully compare results with state-of-the-art in this area is practically impossible because most methods are too complicated and insufficiently described to be implemented in the acceptable time. Because there is no available benchmark fully supporting segmentation method development, we 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 truth which is unfeasible for natural images. Thus, inevitably all such image sets such as the Berkeley benchmark [8] share the same drawbacks - subjectively generated ground truth regions and limited extent which is very difficult and expensive to enlarge. These problems motivated our preference for random mosaics with randomly filled textures even if they only approximate natural image scenes. The profitable 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.

The segmentation results can be judged [15] either by using manually segmented images as reference [7], or visually by comparing to the original images [12], or just by applying quality measures corresponding to human intuition [7, 12]. However it is difficult to avoid subjective ranking conclusions by using either of above approaches on limited test databases. A prior work on the segmentation benchmark is the Berkeley benchmark [8]. This benchmark contains 300 manually segmented Corel database natural images in its public version. The Berkeley benchmark suffers with subjective ground truth and not ideal consistency error performance criteria, which tolerate unreasonable refinement of the ground truth. Over-segmented machine segmentations have always zero consistency error, i.e., they wrongly suggest an ideal segmentation. The benchmark comparison is based on region borders hence different border localization from the human based drawing can handicap otherwise correct scene segmentation. Another segmentation benchmark Minerva [14] contains 448 colour and grey scale images of natural scenes which are segmented using four different seg-

978-1-4244-2175-6/08/$25.00 ©2008 IEEE

menters, segmented regions are manually labelled and different textural features can be learned from these regions and subsequently used by the kNN supervised classifier. This approach suffers from erroneous ground truth resulting from an imperfect segmenter, manual labelling and inadequate textural feature learning from small regions. Outex Texture Database [11] provides a public repository for three types of empirical texture evaluation test suites. It contains 14 classification test suites, while 1 unsupervised segmentation test set is formed by 100 texture mosaics all using the same regular ground truth template and finally one texture retrieval test set. The test suites are publicly available on the website (http://www.outex.oulu.fi), which allows searching, browsing and downloading of the test image databases. Outex currently provides limited test repository but does not allow results evaluation or algorithms ranking. A psycho-visual evaluation of segmentation algorithms using human observers was proposed in [13]. The test was designed to visually compare two segmentations in each step and to answer if any consensus of the best segmentation exists. While such human judgement certainly allows meaningful evaluation, this approach is too demanding to be applicable in image segmentation research.

2. Benchmark

The Prague texture segmentation data-generator and benchmark is web based (**http://mosaic.utia.cas.cz**) service. The goal of the benchmark is to produce score, performance and quality measures for an algorithm's performance for two main reasons:

1. Compare different algorithms to each other,
2. Track and measure the progress toward human-level segmentation performance over time.

A good experimental evaluation should allow comparison of the current algorithm to several leading alternative algorithms, using as many test images as possible and employing several evaluation measures for comparison (in the absence of one clearly optimal measure). Our benchmark possesses all these features. Single textures as well as the mosaics generation approach were chosen on purpose to produce unusually difficult tests to allow an improvement space for future better segmentation algorithms. The benchmark operates either in full mode for registered users (unrestricted mode - U) or in a restricted mode. The major differences between both working modes are that the restricted operational mode does not permanently store visitor's data (results, algorithm details, etc.) into its online database and does not allow custom mosaics creation. To be able to use full-unrestricted benchmark functionalities the user is required to be registered (registration page). The benchmark allows: to obtain customized experimental texture mosaics and their corresponding ground truth (U); to obtain the benchmark texture mosaic sets with their corresponding ground truth; to evaluate visitor's working segmentation results and compare them with state-of-the-art algorithms; to update the benchmark database (U) with an algorithm (reference, abstract, benchmark results) and use it for subsequent other algorithms benchmarking; to grade noise endurance of an algorithm; to check single mosaics evaluation details (criteria values and resulted thematic maps); to rank segmentation algorithms according to the most common benchmark criteria; to obtain LaTeX or MATLAB coded resulting criteria tables (U).

2.1. Image Database

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) rotation invariant texture set, (5) scale invariant texture set, (6) illumination invariant texture set and several invariant combinations (rotation & scale, rotation & illumination, scale & illumination, rotation & scale & illumination). The benchmark uses colour textures from our large (more than 1000 high resolution colour textures categorized into 10 thematic classes) Prague colour texture database. Hard real textures (natural or man-made] were deliberately chosen rather than homogeneous synthesized (for example using Markov RF models) ones because they are more difficult to be correctly segmented for segmentation methods. The benchmark uses cut-outs from the original textures (1/6 approximately). The remaining texture parts are used for the separate test/training sets in the benchmark-supervised mode. These textures were selected deliberately to be difficult for the segmenters. We believe that only under difficult conditions we can obtain useful knowledge for segmentation algorithms improvement.

2.2. Benchmark Generation

Benchmark datasets are computer generated 512×512 random mosaics filled with randomly selected textures. The random mosaics are generated by using a Voronoi polygon random generator. We exploit the fact that segmenting smaller and irregular objects is more difficult than segmenting bigger and regular ob-

jects such as squares or circles. Colour, greyscale or BTF benchmarks are generated upon request in three quantities (20, 80, 180 test mosaics). But if required, it is easy to automatically generate any number of such mosaics (e.g. hundreds or even thousands). For each texture mosaic there are the corresponding ground truth and mask images included. The benchmark enables to test the noise robustness of single segmenters. The benchmark mosaics can be corrupted during their generation with additive Gaussian noise in several signal to noise ratio (SNR) steps, Poisson or salt & pepper noise.

3. Performance Criteria

The submitted benchmark results are evaluated and stored (U) in the server database and used for the algorithm ranking according to a chosen criterion. We have implemented twenty seven most frequented evaluation criteria categorized into four groups: region-based (5+5), pixel-wise (12), 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, over-segmentation, under-segmentation, missed and noise. All these criteria are available either for a single threshold parameter setting or as the performance curves and their integrals. Our pixel-wise criteria group contains the most frequented classification criteria such as the omission and commission errors, class accuracy, recall, precision, mapping score, etc. The consistency criteria group incorporates the global and local consistency errors. Finally, the last criterion set contains three clustering comparison measures. By clicking on a required criterion the evaluation table is reordered, according to this chosen criterion.

4. Examples

The benchmark performance is demonstrated by comparing five unsupervised segmentation algorithms - two our previously published methods GMRF-GM [5] and AR3D-GM [4] and three frequently cited methods JSEG [3], EDISON [2] and Blobworld [1]. The performance of some other methods can be found on the benchmark server. Fig. 1 shows three selected 512×512 mosaics from the colour benchmark created from five to eleven natural colour textures. The last four columns demonstrate comparative results from four alternative algorithms - AR3D-GM, GMRF-GM, Blobword and Edison. Visual comparison suggests over-segmentation inclination of Edison and large missed and noise errors of Blobword. JSEG (not shown here)

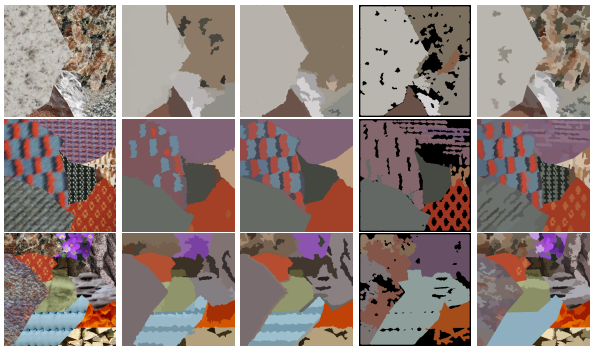

Figure 1. Selected benchmark texture mosaics (a), AR3D-GM (b), GMRF-GM (c), Blobworld (d), and Edison segmentation results (e), respectively.

indicates the second worst both missed and noise errors. The AR3D-GM, GMRF-GM methods produce similar results and both outperform the remaining alternative methods. Integrated numerical results over the whole normal colour benchmark (20 different mosaics) in Tab.1 ($\uparrow$ / $\downarrow$ denote required criterion increase or decrease) confirm these observations. AR3D-GM produces the best correct segmentation, followed by GMRF-GM. JSEG is the third best while Edison is the worst. Edison has strong overegmentation tendency though low ME,NE errors confirm the best inter-region border localization of this method. AR3D-GM and GMRF-GM have slightly less precisely located borders, JSEG doubles these errors and Blobword is by far the worst in this criterion. The pixel-wise criteria (omission error, recall, etc.) further assure the superiority of both AR3D-GM, GMRF-GM methods. Edison leads with small ratio of wrongly assigned pixels (II error) and in the both precision and RM criteria. The consistency criteria confirm their dubiousness. They prefer the Edison method not because of its good performance but due to its high over-segmentation error.

5. Conclusions

The implemented supervised / unsupervised segmentation benchmark is fully automatic web application which enables to mutually compare image segmentation algorithms and to assist in developing new segmentation methods. Segmenters can be ranked based on a chosen criterion from the set of twenty seven different criteria. The test mosaics as well as the ground truths are computer generated which guarantees the evaluation

	Benchmark – Colour				
	AR3D-GM	GMRF-GM	JSEG	Blobworld	EDISON
↑ *CS*	**37.42**	31.93	27.47	21.01	12.68
↓ *OS*	59.53	53.27	38.62	**7.33**	86.91
↓ *US*	8.86	11.24	5.04	9.30	**0.00**
↓ *ME*	12.55	14.97	35.00	59.55	**2.48**
↓ *NE*	13.14	16.91	35.50	61.68	**4.68**
↓ *O*	**34.32**	33.61	37.94	41.45	73.17
↓ *C*	100.00	100.00	92.77	**58.94**	100.00
↑ *CA*	**59.46**	57.91	55.29	46.23	31.19
↑ *CO*	**64.81**	63.51	61.81	56.04	31.55
↑ *CC*	91.79	89.26	87.70	73.62	**98.09**
↓ *I.*	**35.19**	36.49	38.19	43.96	68.45
↓ *II.*	3.39	3.14	3.66	6.72	**0.24**
↑ *EA*	**69.60**	68.41	66.74	58.37	41.29
↑ *MS*	**58.89**	57.42	55.14	40.36	31.13
↓ *RM*	4.88	4.86	4.96	7.96	**3.21**
↑ *CI*	**73.15**	71.80	70.27	61.31	50.29
↓ *GCE*	12.13	16.03	18.45	31.16	**3.54**
↓ *LCE*	6.69	7.31	11.64	23.19	**3.44**
↓ *dM*	15.43	15.27	**15.19**	20.03	16.84
↓ *dD*	**19.76**	20.63	23.38	31.11	35.37
↓ *dVI*	17.10	17.32	17.37	**15.84**	25.65
↑ $\overline{CS}$	**34.68**	31.04	29.13	19.10	12.95
↓ $\overline{OS}$	53.32	49.74	37.70	**10.81**	76.35
↓ $\overline{US}$	9.24	11.33	6.38	8.35	**0.00**
↓ $\overline{ME}$	19.90	21.92	34.72	58.54	**13.91**
↓ $\overline{NE}$	20.80	23.59	35.38	61.24	**15.29**
↑ $\overline{F}$	**72.08**	70.79	69.23	60.46	47.42

Table 1. Different benchmark criteria (see details in http://mosaic.utia.cas.cz).

objectivity and allows easy generation of extensive test sets which are otherwise infeasible to arrange.

The benchmark enables to test single algorithms on monospectral, multispectral or BTF texture data and to test their noise robustness. Further on, it is possible to test scale, rotation and illumination algorithm invariance or any combination of these properties, so that the researchers can quickly and effectively compare their novel algorithms and verify their performance characteristics. Although the benchmark is primarily designed for texture segmenters it gives also good performance insight for any tested image segmenter. The evaluation part of the benchmark can be modified to use also user defined ground truth, for example hand segmented natural images. Other possible applications such as machine learning, feature selection, image compression, QBIC methods evaluation and some others can easily benefit from the benchmark services as well.

Acknowledgments

This research was supported by the projects 102/08/0593, 1ET400750407, 1M0572, 2C06019, 102/07/1594.

References

[1] S. Belongie, C. Carson, H. Greenspan, and J. Malik. Color and texture-based image segmentation using em and its application to content-based image retrieval. In *ICCV*. IEEE, 1998.

[2] C. Christoudias, B. Georgescu, and P. Meer. Synergism in low level vision. In R. Kasturi, D. Laurendeau, and C. Suen, editors, *Proceedings of the 16th ICPR*, volume 4, pages 150–155, Los Alamitos, August 2002. IEEE Computer Society.

[3] Y. Deng and B. Manjunath. Unsupervised segmentation of color-texture regions in images and video. *IEEE Tr. PAMI*, 23(8):800–810, August 2001.

[4] M. Haindl and S. Mikeš. Colour texture segmentation using modelling approach. *Lecture Notes in Computer Science*, (3687):484–491, 2005.

[5] M. Haindl and S. Mikeš. Model-based texture segmentation. *LNCS*, (3212):306 – 313, 2004.

[6] A. Hoover, G. Jean-Baptiste, X. Jiang, P. J. Flynn, H. Bunke, D. B. Goldgof, K. Bowyer, D. W. Eggert, A. Fitzgibbon, and R. B. Fisher. An experimental comparison of range image segmentation algorithms. *IEEE Tr. PAMI*, 18(7):673–689, July 1996.

[7] S. U. Lee, S. Y. Chung, and R. H. Park. A comparitive performance study of several global thresholding techniques for segmentation. *Computer Vision, Graphics, and Image Processing*, 52:171–190, 1990.

[8] 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. Computer Vision*, volume 2, pages 416–423, July 2001.

[9] M. Meila. Comparing clusterings – an axiomatic view. In *ICML*, pages 577 – 584, 2005.

[10] J. Munkres. Algorithms for the assignment and transportation problems. *Journal of the Society for Industrial and Applied Mathematics*, 5(1):32–38, Mar. 1957.

[11] T. Ojala, T. Maenpaa, M. Pietikainen, J. Viertola, J. Kyllonen, and S. Huovinen. Outex: New framework for empirical evaluation of texture analysis algorithms. In *ICPR*, pages I: 701–706, 2002.

[12] N. Pal and S. Pal. A review on image segmentation techniques. *Pattern Recognition*, 26(9):1277–1294, 1993.

[13] C. W. Shaffrey, I. H. Jermyn, and N. G. Kingsbury. Psychovisual evaluation of image segmentation algorithms. In *ACIVS 2002*, September 2002.

[14] M. Sharma and S. Singh. Minerva scene analysis benchmark. In *Seventh Australian and New Zealand Intelligent Information Systems Conference*, pages 231–235. IEEE, November 2001.

[15] Y. J. Zhang. Evaluation and comparison of different segmentation algorithms. *Pattern Recognition Letters*, 18:963–974, 1997.