Texture Segmentation Benchmark

Stanislav Mikeš and Michal Haindl, *Senior Member, IEEE*

Abstract—The Prague texture segmentation data-generator and benchmark (mosaic.utia.cas.cz) is a web-based service designed to mutually compare and rank (recently nearly 200) different static and dynamic texture and image segmenters, to find optimal parametrization of a segmenter and support the development of new segmentation and classification methods. The benchmark verifies segmenter performance characteristics on potentially unlimited monospectral, multispectral, satellite, and bidirectional texture function (BTF) data using an extensive set of over forty prevalent criteria. It also enables us to test for noise robustness and scale, rotation, or illumination invariance. It can be used in other applications, such as feature selection, image compression, query by pictorial example, etc. The benchmark's functionalities are demonstrated in evaluating several examples of leading previously published unsupervised and supervised image segmentation algorithms. However, they are used to illustrate the benchmark functionality and not review the recent image segmentation state-of-the-art.

Index Terms—Benchmark, image segmentation, texture segmentation, (Un)supervised segmentation, segmentation criteria, scale, rotation and illumination invariants

1 Introduction

Unsupervised or supervised texture segmentation is the prerequisite for successful content-based image retrieval, scene analysis, automatic acquisition of virtual models, image interpretation, quality control, security, medical, and many other applications. Although more than a thousand different methods have already been published [1], [8], [23], [24], [48], [57], [60], [70], [72], this ill-defined problem is still far from being solved; it cannot even be solved in its full generality. In addition to that, very little is known about the properties and behavior of already published segmentation methods and their potential user is left to select from among them randomly due to the absence of infallible counseling. This problem is, among others, due to the missing reliable performance comparison between different techniques because the insufficient effort has been given to developing suitable quantitative measures of segmentation quality that could be used to evaluate and compare segmentation algorithms. Rather than advancing the most promising image segmentation approaches, researchers suggesting novel algorithms are often satisfied with them just being sufficiently different from the previously published ones. Furthermore, the methods are tested on only a few carefully selected positive examples. The similarly tricky problem is that of finding optimal parametrization for a given method, especially for those segmenters that have tens of parameters to be set. The most common method for evaluating the effectiveness of a segmentation method is still subjective evaluation [82], where humans visually compare segmentation results for the tested segmenters on some popular datasets, e.g., [10], [12], [19], [47], [86]. However, many datasets have a limited number of classes and objects per image, e.g., [12], [19], they have less than three instances and two categories per image on average. Such tedious and expensive evaluation is inherently restricted to a small number of predetermined test images and has a very limited and questionable generalizing value. The optimal alternative, namely, checking several variants of a developed method on a large number of test images and carefully comparing the results with the state-of-the-art in this area is practically impossible because most methods are too complicated and insufficiently described to be implemented in an acceptable amount of time. Although no theoretical property of a method can be proven experimentally, such an experimental set can indicate its performance and ranking in comparison with alternative algorithms. Because there is no available benchmark fully supporting segmentation method development, we implemented a solution in the form of a 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 [49] and several other proposed approaches [11], [18], [61], [75], [78], [82], [85], share the same drawbacks – subjectively and costly generated ground truth regions and a limited extent, which is very difficult and expensive to enlarge.

These problems motivate our preference for random mosaics with randomly filled textures even if they only approximate natural image scenes. A profitable feature of this trade-off is the unlimited number of different test images with the corresponding objective and the free ground truth map available for each of them.

Zhang [84] differentiates between two types of segmentation evaluation – inter-techniques for ranking the performance of different techniques in segmenting the same type of images, and intra-techniques for recognizing the behavior of the considered technique in segmenting various kinds of

• *The authors are with the Institute of Information Theory and Automation, Czech Academy of Sciences, 11720 Prague, Czech Republic. E-mail: {xaos, haindl}@utia.cas.cz.*

Manuscript received 5 October 2020; revised 12 February 2021; accepted 22 April 2021. Date of publication 27 April 2021; date of current version 4 August 2022.
(Corresponding author: Michal Haindl.)
Recommended for acceptance by M. Varma.
Digital Object Identifier no. 10.1109/TPAMI.2021.3075916

images. Our benchmark is capable of supporting both of these evaluation types.

The segmentation results can be judged [82] either by using manually segmented images as reference (discrepancy methods) [45] possibly with the help of some annotating tool ([16], or by visually comparing them with the original images [57], or just by applying quality measures corresponding to human intuition (goodness methods, unsupervised evaluation) [6], [7], [11], [45], [46], [57], [65], [82], [83]. However, it is challenging to propose general-purpose goodness criteria and to avoid subjective ranking conclusions by using any of the approaches mentioned above on limited test databases. Goodness criteria often suffer from artificial assumptions, e.g., simple, not ragged boundaries, simple interiors [33], intra-region grey-level uniformity [46], or with improper behavior, such as bias towards meaningless single region segmentation [83], etc. The authors [82] have established that a majority of their test goodness measures strongly favored machine segmentations over human segmentations.

Prior work on the segmentation benchmark is the Berkeley benchmark (BSDS300) presented by Martin *et al.* [49]. This benchmark contains more than a thousand various natural images (300 in its public version) from the Corel database, each of which is manually processed by a group of people to get the ground-truth segmentation in the form of the partitioning of the image into a set of disjoint segments. Without any specific guidance, such manual segmentations reflect the subjective human perception, and therefore, different people usually construct different ground truths on the same image. The Berkeley benchmark suffers from several drawbacks. Apart from the subjective ground truth, its performance criteria, i.e., global consistency error (GCE) and local consistency error (LCE), tolerate the ground truth's unreasonable refinement. Over-segmented machine segmentations always have 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. The enlarged version of this benchmark (BSDS500, [3]) uses the original BSDS300 database for training and novel 200 images for testing.

Another segmentation benchmark, Minerva [69], contains 448 color and greyscale images of natural scenes. They are segmented using four different segmenters, the segmented regions are manually labeled 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 labeling, and inadequate textural feature learning from small regions.

Outex Texture Database [56] provides a public repository for three types of empirical texture evaluation test suites. It contains 14 classification test suites, one unsupervised segmentation test suite, which is formed by 100 texture mosaics, and finally, one texture retrieval test suite. All mosaics use the same simple regular ground truth template. The test suites are publicly available on the website (www.outex.oulu.fi), which allows for searching, browsing and downloading the test image databases. Outex currently provides a limited test repository but does not allow for result evaluation or ranking of single algorithms.

A psycho-visual evaluation of segmentation algorithms using human observers was proposed in [68]. The test was designed to visually compare two segmentations in each step and answer whether the best segmentation consensus exists. While such human judgment indeed allows for meaningful evaluation, this approach is too demanding to be applicable in image segmentation research.

The next section describes the basic functionality of our benchmark, the data used, and the benchmark generation algorithm. The following sections present the benchmark performance criteria (3), ranking stability (4), verification on real images (5), examples of different benchmark data with detailed evaluation of ten recently published segmentation methods (6), and conclusions (7).

2 Benchmark

The Prague texture segmentation data-generator and benchmark has been a web-based (mosaic.utia.cas.cz) service already for fourteen years. Although the benchmark has continuously and significantly been upgraded and new features have been appended during this period, it has maintained its backward compatibility and various segmenters tested during these years can still be mutually compared. The goal of the benchmark is to produce a score, performance, and quality measures for an algorithm's performance for two main reasons: different algorithms can be compared to each other, and the progress toward human-level segmentation performance can be tracked and measured over time. A good experimental evaluation should allow for comparison of the current algorithm with several leading alternative algorithms, using as many test images as possible and employing several evaluation measures for such comparison (in the absence of one clearly optimal measure). Our benchmark possesses all of these features.

Single textures and the mosaics generation approaches have been chosen purposefully, namely, to produce unusually complicated tests to provide a space for future segmentation algorithm improvement.

This benchmark allows us to evaluate numerous segmenter's performance characteristics on a virtually unlimited extent of data. However, the number of tested features requires careful consideration to include only the most important ones. Otherwise, the evaluation tables would be fragmented into many specialized sub-tables with few comparative results, and the benchmark would lose its chief purpose.

All test regions are created from naturally measured textures (stochastic, regular, and near-regular, indoor or outdoor); hence they obey the fundamental texture property – homogeneity at least to a certain degree. Textures may limit the validity of the evaluation results on entirely different (textureless) visual data types, for example, segmentation of drawings, cartoons, cartographic maps, documents, range maps, characters, or 3D scenes with significant geometric distortions. Luckily, most existing images, such as outdoor or indoor photographs, aerial or satellite images [28], [67], material samples [27], [31] or medical images [32] are reasonably approximated by these mosaics, and the benchmark ascertainments are informative for them as well.

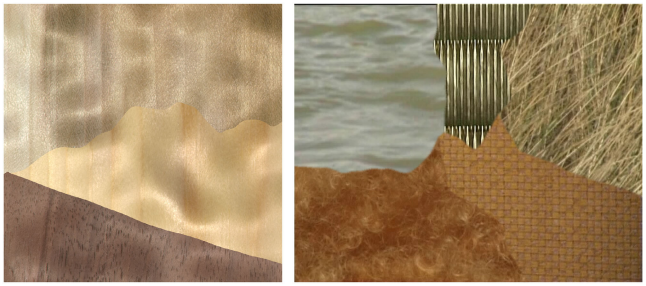

Fig. 1. Sample mosaics with BTF (left) and dynamic (right) textures.

Our benchmark operates either in the 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 data (results, algorithm details, etc.) into its online database and does not allow custom mosaic creation. To be able to use the full-unrestricted benchmark functionalities, the user is required to be registered (on the registration page).

The benchmark allows:

- To obtain customized experimental texture mosaics and their corresponding ground truths (U);
- To obtain the comparative benchmark texture mosaic sets with their corresponding ground truths;
- To evaluate visitor's working segmentation results and compare them with the state-of-the-art algorithms;
- To update the benchmark database (U) with an algorithm (reference, abstract, benchmark results, and code) and use it for subsequent benchmarking of other algorithms;
- To grade the noise, scale, rotation, spectral, illumination, border localization, or benchmark size endurance of an algorithm;
- To check single mosaic evaluation details (criteria values and the resulting thematic maps);
- To rank segmentation algorithms according to the most common benchmark criteria;
- To rank segmentation algorithms according to the weighted meta-criteria of any criteria subset (U);
- To obtain LaTeX or MATLAB coded resulting criteria tables (U) and sensitivity graphs.

2.1 Image Database

The generated texture mosaics, as well as the benchmarks, are composed of the following texture types:

1) Color textures.
2) Monospectral textures (derived from the corresponding color textures).
3) Dynamic color textures (Fig. 1 – right).
4) Hyperspectral (10 spectral bands, 30 m resolution) satellite textures.
5) High-resolution (up to 0.41 m) three-spectral satellite textures.
6) BTF (bidirectional texture function) textures (variable light and viewing angles; variable light and fixed viewing angles; fixed light and variable viewing angles; BTF mapped on variable surface mosaic Fig. 1 – left).
7) Rotation invariant texture set.
8) Scale-invariant texture set.
9) Illumination invariant texture set.
10) Several invariant combinations (rotation & scale, rotation & illumination, scale & illumination, rotation & scale & illumination).

It is thus possible to evaluate how a segmenter's performance depends on the texture scale, illumination and viewing angles, spectral bands, resolution, time, noise, border type, or rotation. The benchmark uses color textures from our large (more than 2000 high resolution color textures categorized into 14 thematic classes and 20 subclasses) Prague color texture database. All these textures are natural textures or man-made material textures, which are only approximately homogeneous (i.e., the local statistics for single textures are similar but not identical). Hard natural textures were deliberately chosen rather than homogeneous synthesized (for example, using Markov random field models) ones because they are significantly more difficult to be correctly segmented by segmentation methods. The benchmark uses cut-outs from the original textures (1/6 approximately), either in the original resolution or a sub-sampled version. The remaining texture parts are used for the separate test/training sets in the benchmark-supervised mode. The benchmarks use 114 color/grey-scale textures from 10 classes. These textures were selected deliberately to be difficult for the segmenters. We believe that only under challenging conditions we can obtain knowledge useful for improving segmentation algorithms. The benchmark has a large margin for improvement for unsupervised segmenters even after fourteen years of service to the community. However, approaching the benchmark limits, a more complex new texture set can be easily introduced without influencing the benchmark concept or its implementation. The BTF measurements [66] are provided courtesy of the University of Bonn, or they are from our UTIA BTF database [26]. Dynamic textures are from the DynTex database [62] and the remote sensing data are either from hyperspectral ALI EO-1 [17] or very-high-resolution GeoEye [25] satellites.

2.2 Benchmark Generation

Benchmark datasets are computer-generated 512×512 (256×256 – Bonn BTF, 1024×1024 – UTIA BTF, 720×576 – dynamic textures) pixel random mosaics filled with randomly selected textures. The random mosaics are generated by using the Voronoi polygon random generator [71]. It first creates a Delaunay triangulation, then it determines the circumscribe centers of its triangles, and finally, it interconnects these points according to the neighborhood relations between the triangles. The resulting Voronoi polygons can further be modified, if required, by inserting additional border points into each polygon line. Alternatively, to piece-wise linear borders, it is possible to generate spline defined borders or suppressed borders using border area morphing. We make use of the fact that segmenting smaller, unequal size and irregular objects is more difficult than segmenting bigger and regular objects, such as obligatory squares or circles. BTF mosaics are created by BTF wood species measurements mapped on 3D spline surfaces fitted into random height lattice coordinates. Dynamic texture mosaics have varying layouts, and each variable cell is filled with a dynamic color texture from the Dyntex database [62]. The

layout is generated from randomly placed control points, which are subsequently randomly shifted between key-frames and interpolated for five frames between two successive key-frames. The area and location of each class region are thus dynamically changed. Invariant mosaics are generated by applying the corresponding transformations, i.e., rotation, scaling, illumination, and variable lighting or viewing angle.

Color and greyscale benchmarks are generated upon request in three quantities (normal = 20, large = 80, huge = 180 test mosaics). BTF and satellite benchmarks have halved quantities. But if required, it is easy to generate any number automatically of such mosaics (e.g., hundreds or even thousands). The benchmark archive, either in the compressed tar or in zip formats, contains images in the PNG format and the data.xml file with detailed descriptions of all mosaics (number of regions, source component textures, size, etc.). For each texture mosaic, the corresponding ground truth and mask images are also included. The supervised version of each benchmark additionally contains independent (i.e., holdout test estimate) training images for every texture class involved. The test mosaic layouts and each cell texture membership are pseudo-randomly generated but with identical initialization of the corresponding random generators, so the requested benchmark sets (for the same size and type) are identical for every visitor.

2.2.1 Noise Corruption

Noise is an important attribute that affects the performance of learning or segmenting algorithms. In real-world applications, noise is an integral part of measurements, and the noise level is usually unknown. The benchmark enables us 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. The user can choose between ten SNR steps for the additive Gaussian ($\langle -10; 35\rangle\ dB$) noise or ten steps for the salt & pepper noise (noise probabilities $\langle 0.5; 0.01\rangle$).

2.2.2 Custom Mosaics

Registered users can benefit from all functions of the underlying benchmark engine. They can design their custom mosaics by specifying the image size, number of cells, number, and type of the textures to be used as well as the type of cell borders (straight lines, piecewise linear, splines, or attenuated borders).

2.2.3 Comparative Methods

For each compared algorithm, there is a concise description available. Each method can contain hyperlinks to further information (author, algorithm details, BIB entry, and WWW external page). Working versions of a segmenter can be compared in the restricted mode. Uploaded temporal results and data in this mode are stored in the database for a limited time only, and they are deleted after its expiration.

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 or the weighted meta-criteria. These user-specified ranking weights allow modifying the relative criteria importance. We have implemented the forty-two (up to now) most frequent evaluation criteria categorized into seven groups: region-based (five criteria with the standard threshold + five performance curves Fig. 5 and five performance integrals (1) through all threshold settings Fig. 5, 12 pixel-wise, four consistency measures, five clustering comparison, three information, seven set criteria, and one boundary criterion. The performance criteria mutually compare the ground truth image regions (or another segmentation) with the corresponding machine segmented regions. The implemented criteria differ in their properties. The subset of informative criteria depends on an application, type of data, or properties the user needs to study. Some criteria are highly correlated (JC, DC); thus, their simultaneous usage has no information value. Some implemented criteria have metric properties (DHD, M, VD, VI), some are bounded mostly between 0 and 1 (BCE, BGM, C, CA, CC, CI, CO, DC, EA, F, FMI, GBCE, GCE, I, II, JC, L, LCE, NMI, O, RI, NBDE) or from one side only (ARI, RM). Ideally, the criteria should be monotonic, symmetric (BCE, BGM, DC, FMI, GBCE, GCE, JC, LCE, M, MI, NMI, VD, VI), independent of the number of pixels (AVI) or segments, and applicable for both supervised or unsupervised segmentation regardless of the number of segments. Segmentation performance can be alternatively or additionally based on boundary matching [23], [36], but these measures are intolerant to sampling, scaling, compression, over-segmentation, and only loosely correspond to human perception of segmentation quality. The benchmark allows us to evaluate and compare single algorithms border precision using either the NBDE criterion or the correct detection (CS) performance curve (Section 3.1) close to the highest evaluated threshold (Fig. 5; $t = 0.975$) which is a robust criterion for the boundary detection precision.

Symbols $\uparrow$ / $\downarrow$ further denote the trend of the corresponding criterion value for the better segmenter, i.e., values $\uparrow$ higher or $\downarrow$ lower than those achieved by an inferior method. All criteria are available on two levels:

- averaged over the corresponding benchmark;
- averaged over benchmark subsets (mosaics sharing the same generation parameters; for normal size, it is only a single mosaic).

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 frequent 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. The clustering comparison group contains five criteria while the information criteria group has three criteria. Seven criteria are implemented in the set group. Finally, the last criterion set contains the boundary displacement error. The evaluation table is reordered according to the chosen criterion by clicking on a required criterion or the meta-criterion.

3.1 Region-Based Criteria

The region-based criteria [35] mutually compare the machine segmented regions R_i $i = 1, \ldots, M$ with the correct ground truth (or another segmentation) regions $\tilde{R}_j$ $j = 1, \ldots, N$ where $|R|$ is the corresponding set cardinality. The regions overlap acceptance is controlled by the threshold $t = 0.75$. Single region-based criteria are defined as follows:

$\uparrow CS$ (correct detection) : $\{R_k, \tilde{R}_{\hat{k}}\}$ iff

1) $|R_k \cap \tilde{R}_{\hat{k}}| \geq t\,|R_k|$,
2) $|R_k \cap \tilde{R}_{\hat{k}}| \geq t\,|\tilde{R}_{\hat{k}}|$.

The ideal segmentation has the same number of correctly detected (CS) regions with very similar shapes and locations (for the required 75 percent overlap) as the ground truth map. Ideally, neither ground truth region should be over-segmented, nor should any machine segmented region contain more than one corresponding ground truth region (under-segmentation).

$\downarrow OS$ (over-segmentation) : $\{R_{k_1}, \ldots, R_{k_x}, \tilde{R}_{\hat{k}}\}$, $2 \leq x \leq M$ iff

1) $\forall i \in \langle 1; x\rangle, \quad |R_{k_i} \cap \tilde{R}_{\hat{k}}| \geq t\,|R_{k_i}|$,
2) $\sum_{i=1}^{x} |R_{k_i} \cap \tilde{R}_{\hat{k}}| \geq t\,|\tilde{R}_{\hat{k}}|$.

$\downarrow US$ (under-segmentation) : $\{R_k, \tilde{R}_{\hat{k}_1}, \ldots, \tilde{R}_{\hat{k}_x}\}$, $2 \leq x \leq N$ iff

1) $\sum_{i=1}^{x} |R_k \cap \tilde{R}_{\hat{k}_i}| \geq t\,|R_k|$,
2) $\forall i \in \langle 1; x\rangle, \quad |R_k \cap \tilde{R}_{\hat{k}_i}| \geq t\,|\tilde{R}_{\hat{k}_i}|$.

The missed regions are the ground truth regions that have not been detected in any of the categories mentioned above (CS, OS, US).

$\downarrow ME$ (missed error) : $\{\tilde{R}_{\hat{k}}\}$ iff $\tilde{R}_{\hat{k}} \notin CS$, $\tilde{R}_{\hat{k}} \notin OS$, $\tilde{R}_{\hat{k}} \notin US$.

Similarly, the noise regions are the machine segmented regions which do not belong to any of the CS, OS, or US categories.

$\downarrow NE$ (noise error) : $\{R_k\}$ iff $R_k \notin CS$, $R_k \notin OS$, $R_k \notin US$.

Single region-based criteria are also available as the corresponding performance curves, see Fig. 5 $CS(t)$, $OS(t)$, $US(t)$, $ME(t)$, $NE(t)$. These curves allow us to compare sensitivity of different segmenters to the changing threshold value ($t \in \langle 0.5; 1\rangle$). Finally, the last five region criteria are approximations of the performance curve integrals

$$\bar{f} = 2\int_{0.5}^{1} f(t)\,dt\,, \tag{1}$$

where $f(t)$ is a curve from {$CS(t)$, $OS(t)$, $US(t)$, $ME(t)$, $NE(t)$}. These integral criteria can be found in the parentheses (Fig. 5) next to the algorithm color symbol, but not in the results comparison tables' page.

3.2 Pixel-Wise Weighted Average Criteria

The pixel-wise criteria were originally developed for the supervised classifiers evaluation. We also generalized them for the unsupervised (i.e., unknown class-separate training sets and number of classes) applications, where their direct application is prevented due to the unknown mutual correspondence between the segmented and ground truth regions, as well as the different cardinalities of both these region sets. The mutual assignment of the machine segmented and ground truth regions for the pixel-wise criteria evaluation is solved by using the Munkre's assignment algorithm [55] which finds the minimal cost assignment $g : A \mapsto B$, $\sum_{\alpha \in A} f(\alpha, g(\alpha))$ between sets A, B, $|A| = |B| = n$ given by the cost function $f(\alpha, \beta)$, $\alpha \in A, \beta \in B$. The algorithm has polynomial complexity instead of exponential for the exhaustive search.

Let us denote $n_{i,\bullet} = \sum_{j=1}^{N} n_{i,j}$, and $n_{\bullet,i} = \sum_{j=1}^{M} n_{j,i}$, where N, M are the correct number of classes and the interpreted number of classes (or regions), respectively. n is the number of pixels in the test set, $K = \max\{M, N\}$, $n_{i,j}$ is the number of pixels interpreted as the ith class but belonging into the jth class. The error matrix ($\{n_{i,j}\}$) extended into $K \times K$ is obtained by padding missing entries with zeros. Here $\hat{\imath}$ is either i for supervised tests or mapping of the ith class ground truth into an interpretation segment based on the Munkres algorithm (for an unsupervised test). The following pixel-wise criteria were implemented:

The overall ratio of the wrongly interpreted pixels $\downarrow O$ (omission error)

$$O = \operatorname{med}\left\{\frac{O_i}{n_{\bullet,i}}\right\}_{i=1}^{N} = \operatorname{med}\left\{1 - \frac{n_{\hat{\imath},i}}{n_{\bullet,i}}\right\}_{i=1}^{N} \langle 0; 1\rangle\,,$$

where O_i is the ith class omission error. The overall ratio of the wrongly assigned pixels $\downarrow C$ (commission error)

$$C = \operatorname{med}\left\{\frac{C_i}{n_{\hat{\imath},\bullet}}\right\}_{\hat{\imath}=1}^{M} = \operatorname{med}\left\{1 - \frac{n_{\hat{\imath},i}}{n_{\hat{\imath},\bullet}}\right\}_{\hat{\imath}=1}^{M} \langle 0; 1\rangle\,,$$

where C_i is the ith class commission error.

$\uparrow CA$ (the weighted average class accuracy)

$$CA = \frac{1}{n}\sum_{i=1}^{K} \frac{n_{\hat{\imath},i}\, n_{\bullet,i}}{n_{\bullet,i} + n_{\hat{\imath},\bullet} - n_{\hat{\imath},i}} \langle 0; 1\rangle\,,$$

$\uparrow CO$ (recall, the weighted average correct assignment)

$$CO = \frac{1}{n}\sum_{i=1}^{K} n_{\bullet,i}\, CO_i = \frac{1}{n}\sum_{i=1}^{K} n_{\hat{\imath},i} \langle 0; 1\rangle\,,$$

$\uparrow CC$ (precision, object accuracy, overall accuracy)

$$CC = \frac{1}{n}\sum_{i=1}^{K} n_{\bullet,i}\, CC_i = \frac{1}{n}\sum_{i=1}^{K} \frac{n_{\hat{\imath},i}\, n_{\bullet,i}}{n_{\hat{\imath},\bullet}} \langle 0; 1\rangle\,,$$

$\downarrow I.$ (type I error, the weighted probability of wrong assignment of classes pixels)

$$I = \frac{1}{n}\sum_{i=1}^{K} \left(n_{\bullet,i} - n_{\hat{\imath},i}\right) = 1 - CO \langle 0; 1\rangle\,,$$

$\downarrow II.$ (type II error, the weighted probability of commission error)

$$II = \frac{1}{n}\sum_{i=1}^{K} \frac{n_{\hat{\imath},\bullet}\, n_{\bullet,i} - n_{\hat{\imath},i}\, n_{\bullet,i}}{n - n_{\bullet,i}} \langle 0; 1\rangle\,,$$

$\uparrow EA$ (mean class accuracy estimate)

$$EA = \frac{1}{n}\sum_{i=1}^{K} \frac{2 n_{\hat{\imath},i}\, n_{\bullet,i}}{n_{\bullet,i} + n_{\hat{\imath},\bullet}} \langle 0; 1\rangle\,.$$

The $\uparrow F$ measure curve

$$F(\gamma) = \frac{1}{n}\sum_{i=1}^{K} n_{\bullet,i}\,\frac{CC_i\,CO_i}{\gamma CO_i + (1-\gamma)CC_i}\langle 0;1\rangle\,,$$

where $\gamma \in \langle 0;1\rangle$. $F(0.5) = EA$, $F(0) = CO$, $F(1) = CC$.

The mapping score $\uparrow MS$ emphasizes the error of not recognizing the test data

$$MS = \frac{1}{n}\sum_{i=1}^{K}\big(1.5\,n_{i,i} - 0.5 n_{\hat{i},\bullet}\big)\langle -0.5;1\rangle\,.$$

The root mean square proportion estimation error $\downarrow RM$

$$RM = \sqrt{\frac{1}{K}\sum_{i=1}^{K}\Big(\frac{n_{\hat{i},\bullet} - n_{\bullet,i}}{n}\Big)^2} \geq 0$$

indicates an unbalance between the omission O_i and commission C_i errors, respectively. The comparison index $\uparrow CI$ includes both these types of errors

$$CI = \frac{1}{n}\sum_{i=1}^{K} n_{i,i}\sqrt{\frac{n_{\bullet,i}}{n_{\hat{i},\bullet}}} = \frac{1}{n}\sum_{i=1}^{K} n_{\bullet,i}\sqrt{CC_i CO_i}\ \langle 0;1\rangle\,,$$

where CC_i, CO_i are the object precision and recall. CI reaches its maximum either for the ideal segmentation or for equal commission and omission errors for every region (class).

3.3 Consistency Error Criteria

Let $S, \tilde{S}$ be two segmentations, R_p the set of pixels corresponding to a region in the S segmentation and containing the pixel p, $|R|$ the set cardinality and $\backslash$ the set difference. A refinement tolerant measure error was defined [49] at each pixel p

$$\varepsilon_p(S,\tilde{S}) = \frac{|R_p \backslash \tilde{R}_p|}{|R_p|}\,.$$

This non-symmetric local error measure encodes a measure of refinement in one direction only. Two symmetric error measures for the entire image, based on the theory of the human perceptual organization, are defined: Global Consistency Error ($\downarrow GCE$) forces all local refinements to be in the same direction while Local Consistency Error ($\downarrow LCE$) allows refinement in both directions

$$GCE(S,\tilde{S}) = \frac{1}{n}\min\left\{\sum_p \varepsilon_p(S,\tilde{S}), \sum_p \varepsilon_p(\tilde{S},S)\right\},$$

$$LCE(S,\tilde{S}) = \frac{1}{n}\sum_p \min\big\{\varepsilon_p(S,\tilde{S}), \varepsilon_p(\tilde{S},S)\big\}\,,$$

$$LCE, GCE \in \langle 0;1\rangle\,, \qquad LCE \leq GCE\,.$$

The major problem with these consistency measures is their tolerance to incorrect over-segmentation of the ground truth. If the segmentation is an over-segmented version of the ground truth or vice versa, the segmentation error is always zero. Thus the trivial segmentations with either all regions containing just one pixel or the whole image being a single region are the ideal segmentations $LCE = GCE = 0$ according to both consistency criteria. To overcome this problem, Martin proposed [50] Bidirectional Consistency Error ($\downarrow BCE$) and further Šrubař [73] Global Bidirectional Consistency Error ($\downarrow GBCE$)

$$BCE(S,\tilde{S}) = \frac{1}{n}\sum_p \max\big\{\varepsilon_p(S,\tilde{S}), \varepsilon_p(\tilde{S},S)\big\}\,,$$

$$GBCE(S,\tilde{S}) = \frac{1}{n}\max\left\{\sum_p \varepsilon_p(S,\tilde{S}), \sum_p \varepsilon_p(\tilde{S},S)\right\},$$

$$BCE, GBCE \in \langle 0;1\rangle\,.$$

3.4 Clustering Comparison Criteria

Several clustering comparison criteria – BGM, SC, SSC, VD, L are implemented in the benchmark. Denote $n_{k,\tilde{k}}$ the number of points in the intersection $n_{k,\tilde{k}} = |R_k \cap \tilde{R}_{\tilde{k}}|$, $n_k = |R_k|$, $n_{\tilde{k}} = |\tilde{R}_k|$.

$\uparrow BGM$ (bipartite graph matching [42])

$$BGM(S,\tilde{S}) = \frac{w}{n}\langle 0;1\rangle\,,$$

where w is the sum of a maximum-weight bipartite graph matching. The nodes are clusters of S and $\tilde{S}$, while the edges are between each pair of nodes $(k,\tilde{k})$ having the weight $n_{k,\tilde{k}}$. Similarly $\downarrow VD$ the bounded n-invariant, symmetric Van Dongen metric evaluates only intersections using Directional Hamming Distances $\downarrow DHD$ as defined in [36].

Another method of this group computes set differences. For this purpose, maximal intersections are removed. First, we define Directional Hamming Distance from segmentation S to segmentation $\tilde{S}$ [36]

$$DHD(S,\tilde{S}) = \sum_{\tilde{k}}\left(\sum_k n_{k,\tilde{k}} - \max_k n_{k,\tilde{k}}\right),$$

$$VD(S,\tilde{S}) = \frac{DHD(S,\tilde{S}) + DHD(\tilde{S},S)}{2n}\,,$$

$$VD(S,\tilde{S}) = 1 - \left(\sum_k \max_{\tilde{k}} n_{k,\tilde{k}} - \sum_{\tilde{k}} \max_k n_{k,\tilde{k}}\right)\Big/ 2n\,.$$

$\uparrow L$ (Larsen [44] asymmetric criterion)

$$L(S,\tilde{S}) = \frac{1}{n_k}\sum_k \max_{\tilde{k}}\frac{2n_{k,\tilde{k}}}{n_k + n_{\tilde{k}}}\langle 0;1\rangle\,,$$

two directional segmentation covering $\uparrow SC$

$$SC(S,\tilde{S}) = \frac{1}{n}\sum_k n_k \max_{\tilde{k}}\frac{n_{k,\tilde{k}}}{n_k + n_{\tilde{k}} - n_{k,\tilde{k}}}\langle 0;1\rangle\,,$$

and $\uparrow SSC$

$$SSC(S,\tilde{S}) = SC(\tilde{S},S)\langle 0;1\rangle\,.$$

3.5 Information Criteria

We have implemented three normalized information-based criteria derived from the variation of information VI and the mutual information MI. Let denote the entropy as

$$H(S) = -\sum_k \frac{n_k}{n} \log_2 \frac{n_k}{n}$$

and $\uparrow MI$ (the symmetric mutual information)

$$MI(S,\tilde{S}) = \sum_k \sum_{\tilde{k}} \frac{n_{k,\tilde{k}}}{n} \log_2 \frac{n_{k,\tilde{k}}\, n}{n_k\, n_{\tilde{k}}}$$

is bounded by entropies $0 \le MI(S,\tilde{S}) \le \min\{H(S), H(\tilde{S})\}$.

Then the variation of information $\downarrow VI$ (is a metric) not bounded by a constant value defined [51]:

$$VI(S,\tilde{S}) = H(S) + H(\tilde{S}) - 2MI(S,\tilde{S}) \langle 0; \log_2 n \rangle .$$

It is possible to show that the variation of information complies with symmetry, additivity w.r.t. refinement, additivity w.r.t. join, convex additivity and scale properties (see the details in [51]).

$\downarrow AVI$ (VI normalized w.r.t. number of pixels)

$$AVI(S,\tilde{S}) = \frac{VI}{\log_2(n)} \langle 0; 1 \rangle .$$

$\downarrow NVI$ (VI normalized w.r.t. number of classes/regions)

$$NVI(S,\tilde{S}) = \frac{VI}{2\log_2(\max\{K,\tilde{K}\})} ,$$

where $NVI(S,\tilde{S}) \in \left\langle 0; \frac{\log_2 n}{2\log_2(\max\{K,\tilde{K}\})} \right\rangle$.

$\uparrow NMI$ (normalized mutual information [74])

$$NMI(S,\tilde{S}) = \frac{MI(S,\tilde{S})}{\sqrt{H(S)H(\tilde{S})}} \langle 0; 1 \rangle .$$

3.6 Set Criteria

Set measures are based on counting pixel pairs [42], [76] which are either in the same regions in both partitions S and $\tilde{S}$ – N_{11}, different regions in both partitions – N_{00}, the same regions in S but different regions in $\tilde{S}$ – N_{10}, and different regions in S but the same regions in $\tilde{S}$ – N_{01}. These criteria are symmetric with the exception of WI and WII criteria. Note that the count of pairs in $(N_{11} + N_{00})$ represents the agreement whereas the count of pairs in $(N_{10} + N_{01})$ represents the disagreement between the two partitions.

$$N_{11} = \sum_k \sum_{\tilde{k}} \binom{n_{k,\tilde{k}}}{2}$$

$$N_{10} = \sum_k \binom{n_k}{2} - N_{11}$$

$$N_{01} = \sum_{\tilde{k}} \binom{n_{\tilde{k}}}{2} - N_{11}$$

$$N_{00} = \binom{n}{2} - N_{11} - N_{10} - N_{01}$$

$\uparrow JC$ (Jaccard coefficient [41])

$$JC(S,\tilde{S}) = \frac{N_{11}}{N_{11} + N_{10} + N_{01}} \langle 0; 1 \rangle .$$

$\uparrow DC$ (Dice coefficient [15])

$$DC(S,\tilde{S}) = \frac{N_{11}}{N_{11} + (N_{10} + N_{01})/2} \langle 0; 1 \rangle .$$

The Dice coefficient can be computed from the Jaccard coefficient:

$$DC(S,\tilde{S}) = \frac{2JC(S,\tilde{S})}{1 + JC(S,\tilde{S})} ,$$

$DC(S,\tilde{S}) \le JC(S,\tilde{S})$, thus it provides the identical ranking.

$\uparrow FMI$ (Fowlkes and Mallows index [21])

$$FMI(S,\tilde{S}) = \frac{N_{11}}{\sqrt{(N_{11} + N_{10})(N_{11} + N_{01})}} \langle 0; 1 \rangle .$$

FMI has high value for small number of cluster even for independent partitions.

$\uparrow ARI$ (adjusted RI [38]) is based on Rand index [64]:

$$RI(S,\tilde{S}) = \frac{2(N_{11} + N_{00})}{n(n-1)} \langle 0; 1 \rangle .$$

The Rand index is dependent on the number of clusters and elements. RI converges to 1 as the number of clusters increases in independent clusterings [21] and its adjustement is:

$$ARI = \frac{RI - ExpectedIndex}{MaxIndex - ExpectedIndex} \qquad \le 1.$$

$$ARI = \frac{\sum_{k,\tilde{k}} \binom{n_{k,\tilde{k}}}{2} - \left[\sum_k \binom{n_k}{2} \sum_{\tilde{k}} \binom{n_{\tilde{k}}}{2}\right] / \binom{n}{2}}{\frac{1}{2}\left[\sum_k \binom{n_k}{2} + \sum_{\tilde{k}} \binom{n_{\tilde{k}}}{2}\right] - \left[\sum_k \binom{n_k}{2} \sum_{\tilde{k}} \binom{n_{\tilde{k}}}{2}\right] / \binom{n}{2}}$$

ARI problematic assumptions is a generalized hypergeometric distribution for null hypothesis [38] and possible negative index values [51].

$\uparrow WI, WII$ (Wallace [77])

$$WI(S,\tilde{S}) = \frac{N_{11}}{N_{11} + N_{01}} \langle 0; 1 \rangle ,$$

$$WII(S,\tilde{S}) = \frac{N_{11}}{N_{11} + N_{10}} \langle 0; 1 \rangle .$$

Criteria WI, WII, DC are the same as P_r, R_r, F_r in [63] and symmetric FMI is the geometrical mean of asymetric criteria WI, WII, i.e., $FMI(S,\tilde{S}) = \sqrt{WI(S,\tilde{S})\, WII(S,\tilde{S})}$.

The Mirkin metric $\downarrow M$ is defined:

$$M(S,\tilde{S}) = \frac{2(N_{01} + N_{10})}{n(n-1)} \ge 0 ,$$

$$M(S,\tilde{S}) = 1 - RI(S,\tilde{S}) \Big/ \binom{n}{2}.$$

The Mirkin metric is sensitive to cluster sizes and wrongly prefers clusterings with identical cluster's cardinality [76].

3.7 Boundary Criterion

$\downarrow BDE$ (boundary displacement error) [79] measures the average displacement error of boundary pixels between two segmentation results. Mean absolute surface distance (MASD) measures the average minimum distance between two boundaries. BDE is elsewhere called MASD. We use normalized version of this criterion

$$d(r, B_2) = \min_{s \in B_2} \|r - s\|$$

$$BDE(B_1, B_2) = \frac{1}{2}\left(\sum_{r \in B_1} \frac{d(r, B_2)}{|B_1|} + \sum_{r \in B_2} \frac{d(r, B_1)}{|B_2|}\right),$$

where $d(r, B_2)$ is the euclidean distance of a boundary point $r \in B_1$ to the boundary set B_2.

$\downarrow NBDE$ (normalized boundary displacement error)

$$NBDE(B_1, B_2) = \frac{2}{\max(w, h)} BDE(B_1, B_2)\langle 0; 1\rangle\,,$$

where w, h is the width and the height of the image.

3.8 Meta-Criteria

The meta-criteria serve for fast ranking of segmenters using any subset of the implemented benchmark criteria. This criteria can have either identical influence ($w_c = \frac{1}{|\mathcal{C}|}, RANK^m \in \langle 1; |\mathcal{M}|\rangle$) or their importance can be individually selected. $\mathcal{C}$ is a chosen criteria set, and the criteria trend is $d_c \in \{\uparrow, \downarrow\}$. $\mathcal{M}$ is a given set of methods / results and the user selected criteria weights are w_c, where $W = \sum_{c \in \mathcal{C}} w_c$. All criteria values x_c^m are multiplied by one hundred ($\times 100$), where $m \in \mathcal{M}$, $c \in \mathcal{C}$:

$$\begin{aligned}
X_c &= \{x_c^m \mid m \in \mathcal{M}\},\\
r_c^m &= \operatorname{rank}(x_c^m, X_c),\\
v_c^m &= \begin{cases} x_c^m, & \text{for } d_c = \uparrow\\ 100 - x_c^m, & \text{for } d_c = \downarrow \end{cases},\\
z_c^m &= \frac{x_c^m - \mu_c}{\rho_c},\ \mu_c = E(X_c),\ \rho_c = \sqrt{\operatorname{var}(X_c)},\\
s_c &= \{+1 \text{ for } d_c = \uparrow, -1 \text{ for } d_c = \downarrow\}.
\end{aligned}$$

$\downarrow RANK$ (weighted average of ranks)

$$RANK^m = \frac{1}{W}\sum_{c \in \mathcal{C}} w_c r_c^m$$

$\uparrow AVG$ (weighted average of values)

$$AVG^m = \frac{1}{W}\sum_{c \in \mathcal{C}} w_c v_c^m$$

$\uparrow NORM$ (weighted average of z-scores)

$$NORM^m = \frac{1}{W}\sum_{c \in \mathcal{C}} w_c s_c z_c^m.$$

3.9 Criteria Relationship

A natural question arises with these many evaluation criteria used by different researchers: are they all really needed? An optimal criterion depends on the intended application and varying classification priorities, which is the reason why so many criteria are used. The unsupervised segmenters in Table 1 illustrate this observation: there is no segmenter scoring best for all of the evaluated criteria. Applications that cannot tolerate over-segmentation cannot use consistency measures or under-segmentation. Security applications and defect detectors should, on the other hand, guarantee low under-segmentation; thus the commission error or Van Dongen metric are not the best criteria to consult. Region-based criteria are robust and appropriate for the majority of applications where precise border location is not of primary interest. For this reason, the benchmark does not prefer any criterion. A user can click on any criterion to reorder the evaluation table according to an intended application or a tested performance characteristic or use the meta-criteria over a subset of criteria.

3.9.1 Pearson Correlation

Fig. 2 presents color-coded Pearson correlation analysis[1] for thirty-six segmentation criteria computed for seventeen unsupervised segmentation algorithms, which were evaluated using our 180 color benchmark test mosaics. While strong correlation between I, CO and EA, CA can be expected, a high correlation between ME, NE or CI, MS criteria is less obvious. In this experiment three mutually positively correlated groups of criteria g_1, g_2, g_3 emerged

$$\begin{aligned}
g_1 &= \{CS, EA, CA, CI, MS, CO, FMI,\\
&\qquad JC, DC, SSC, ARI, BGM, SC\}\\
g_2 &= \{GCE, LCE, ME, NE\}\\
g_3 &= \{I, BCE, GBCE, VD,\\
&\qquad AVI, NVI, M, RM, O\}.
\end{aligned}$$

The g_3 group is simultaneously negatively correlated with the group g_1. The lowest mutual correlation with others occurs for the OS over-segmentatiuon criterion. The same analysis on the Berkeley data set [49] confirms g_1, g_2 positive correlation, but only $g_3^{BSDS} = \{I, M\}$ and g_1 negative correlation with $g_4^{BSDS} = \{M, VD, I\}$. Although such analysis suggests that it is sufficient to use only one representative criterion per correlated criterion's group for the concise evaluation of an algorithm, it is not the case. Single criteria correlation depends on data classes, the number of test mosaics, and a specific classifier. Even one of the most stable correlation values between NE and ME weaker for two methods (DBM [30] and VRA [58]). Thus a detailed analysis of a method's properties is required to study all of our criteria.

3.9.2 Spearman Rank Correlation

Criteria relationship was also analyzed using the Spearman rank correlation for all mutual combinations (630) of 36 criteria using the significance level 0.01. The rank correlation

1. For further graphs see mosaic.utia.cas.cz/PAMI.

TABLE 1
Color Benchmark Results for *EWT-FCNT, *FCNT, †FCNT, A3M, PCA-MS, GRPNMF, CMS, LGG, IGMRF, †RS

	Benchmark – Color									
	*EWT-FCNT	*FCNT	†FCNT	A3M	PCA-MS	GRPNMF	CMS	LGG	IGMRF	†RS
↓ *RANK*	**1.00**	2.00	3.11	4.31	5.00	6.06	7.19	8.22	8.83	*9.28*
↑ *AVG*	**98.43**	95.98	89.21	88.21	87.45	84.98	80.21	76.67	75.46	*71.87*
↑ *NORM*	**1.535**	1.246	0.497	0.380	0.292	0.013	−0.509	−0.898	−1.066	*−1.491*
↑ *CS*	**98.45** [1]	96.01 [2]	79.34 [3]	77.73 [4]	72.27 [5]	69.50 [6]	53.73 [7]	*41.42* [10]	49.02 [8]	46.02 [9]
↓ *OS*	**0.00** [1]	1.56 [2]	13.67 [3]	15.92 [7]	*18.33* [10]	16.05 [8]	16.29 [9]	15.04 [5]	15.06 [6]	13.96 [4]
↓ *US*	**0.00** [1]	1.20 [2]	6.25 [3]	6.31 [4]	9.41 [5]	9.90 [6]	11.76 [8]	12.48 [9]	10.82 [7]	*30.01* [10]
↓ *ME*	**0.37** [1]	0.78 [2]	3.80 [3]	3.93 [4]	4.19 [5]	5.74 [6]	18.57 [8]	*27.64* [10]	21.44 [9]	12.01 [7]
↓ *NE*	**0.46** [1]	0.89 [2]	3.80 [3]	3.92 [5]	3.92 [4]	5.89 [6]	18.66 [8]	*26.92* [10]	22.76 [9]	11.77 [7]
↓ *O*	**0.93** [1]	2.71 [2]	6.48 [3]	7.68 [5]	7.25 [4]	10.33 [6]	12.55 [7]	17.80 [8]	22.06 [9]	*35.11* [10]
↓ *C*	**1.05** [1]	2.29 [2]	22.88 [7]	24.24 [8]	6.44 [3]	18.35 [6]	16.85 [5]	15.13 [4]	26.87 [9]	*29.91* [10]
↑ *CA*	**97.67** [1]	93.95 [2]	84.17 [3]	82.80 [4]	81.13 [5]	77.92 [6]	71.58 [7]	66.53 [8]	64.74 [9]	*58.75* [10]
↑ *CO*	**98.78** [1]	96.73 [2]	87.97 [3]	86.89 [4]	85.96 [5]	84.00 [6]	78.99 [7]	75.75 [8]	73.95 [9]	*68.89* [10]
↑ *CC*	**98.81** [1]	97.02 [2]	94.15 [3]	93.65 [4]	91.24 [5]	88.41 [6]	84.91 [7]	82.19 [8]	79.03 [9]	*69.30* [10]
↓ *I*	**1.22** [1]	3.27 [2]	12.03 [3]	13.11 [4]	14.04 [5]	16.00 [6]	21.01 [7]	24.25 [8]	26.05 [9]	*31.11* [10]
↓ *II*	**0.25** [1]	0.68 [2]	1.42 [3]	1.50 [4]	1.59 [5]	2.26 [6]	3.11 [7]	4.17 [9]	3.99 [8]	*8.63* [10]
↑ *EA*	**98.76** [1]	96.68 [2]	88.97 [3]	88.03 [4]	87.08 [5]	84.39 [6]	79.39 [7]	76.10 [8]	73.98 [9]	*65.87* [10]
↑ *MS*	**98.17** [1]	95.10 [2]	85.23 [3]	83.98 [4]	81.84 [5]	78.57 [6]	70.84 [7]	63.63 [8]	60.92 [9]	*55.52* [10]
↓ *RM*	**0.24** [1]	0.86 [2]	3.12 [3]	3.27 [4]	4.45 [6]	4.34 [5]	6.45 [8]	6.72 [9]	6.01 [7]	*10.96* [10]
↑ *CI*	**98.78** [1]	96.77 [2]	89.91 [3]	89.03 [4]	87.81 [5]	85.24 [6]	80.61 [7]	77.48 [8]	75.18 [9]	*67.35* [10]
↓ *GCE*	**2.23** [1]	5.55 [2]	6.46 [3]	7.40 [4]	8.33 [5]	10.61 [6]	16.29 [8]	20.47 [9]	*23.07* [10]	11.23 [7]
↓ *LCE*	**1.68** [1]	3.75 [2]	4.75 [3]	5.62 [5]	5.61 [4]	7.69 [6]	8.88 [8]	11.25 [9]	*15.90* [10]	7.70 [7]
↓ *BCE*	**2.88** [1]	7.69 [2]	17.73 [3]	19.31 [4]	21.00 [5]	24.22 [6]	30.80 [7]	35.25 [8]	36.10 [9]	*40.64* [10]
↓ *GBCE*	**2.33** [1]	5.89 [2]	16.02 [3]	17.53 [4]	18.28 [5]	21.30 [6]	23.38 [7]	26.03 [8]	28.92 [9]	*37.11* [10]
↑ *BGM*	**98.78** [1]	96.81 [2]	87.97 [3]	86.89 [4]	85.96 [5]	84.00 [6]	78.99 [7]	75.75 [8]	73.95 [9]	*68.89* [10]
↑ *SC*	**97.69** [1]	94.16 [2]	84.11 [3]	82.57 [4]	81.30 [5]	78.67 [6]	72.83 [7]	68.47 [8]	67.21 [9]	*63.92* [10]
↑ *SSC*	**97.67** [1]	94.03 [2]	84.85 [3]	83.49 [4]	82.16 [5]	79.21 [6]	73.35 [7]	68.43 [8]	67.44 [9]	*62.40* [10]
↓ *VD*	**1.21** [1]	3.06 [2]	7.79 [3]	8.57 [4]	9.06 [5]	10.80 [6]	14.43 [7]	17.13 [8]	*18.67* [10]	18.52 [9]
↑ *L*	**97.86** [1]	94.69 [2]	88.53 [3]	87.67 [4]	85.54 [5]	83.08 [6]	79.26 [7]	75.42 [8]	72.20 [9]	*66.67* [10]
↓ *AVI*	**1.02** [1]	2.27 [2]	3.93 [3]	4.39 [5]	4.31 [4]	5.39 [6]	5.92 [7]	6.47 [8]	*8.14* [10]	7.96 [9]
↓ *NVI*	**3.33** [1]	7.56 [2]	10.70 [3]	11.86 [4]	13.22 [5]	15.83 [6]	18.40 [7]	21.84 [8]	*27.33* [10]	25.08 [9]
↑ *NMI*	**96.32** [1]	91.69 [2]	86.81 [3]	85.28 [4]	84.78 [5]	81.02 [6]	78.23 [7]	76.07 [8]	70.26 [9]	*61.66* [10]
↓ *M*	**0.74** [1]	1.96 [2]	4.88 [3]	5.30 [4]	5.88 [5]	7.42 [6]	10.09 [7]	11.22 [9]	10.99 [8]	*23.67* [10]
↑ *ARI*	**97.61** [1]	93.84 [2]	84.85 [3]	83.46 [4]	81.62 [5]	77.88 [6]	70.52 [7]	66.22 [8]	66.08 [9]	*56.50* [10]
↑ *JC*	**96.31** [1]	90.74 [2]	80.02 [3]	78.26 [4]	75.30 [5]	71.59 [6]	64.26 [7]	59.42 [8]	59.42 [9]	*55.50* [10]
↑ *DC*	**98.09** [1]	95.10 [2]	87.75 [3]	86.61 [4]	85.25 [5]	82.22 [6]	76.88 [7]	73.37 [8]	73.13 [9]	*68.10* [10]
↑ *FMI*	**98.09** [1]	95.11 [2]	88.28 [3]	87.16 [4]	85.71 [5]	82.96 [6]	77.31 [7]	73.72 [8]	73.46 [9]	*70.97* [10]
↑ *WI*	**98.14** [1]	94.77 [2]	90.69 [3]	89.88 [4]	88.20 [5]	83.80 [6]	77.60 [7]	74.33 [9]	76.08 [8]	*65.41* [10]
↑ *WII*	**98.04** [1]	95.46 [2]	87.01 [3]	85.64 [4]	84.18 [5]	83.76 [7]	77.93 [8]	73.83 [9]	*71.54* [10]	84.01 [6]
↓ *NBDE*	**0.60** [1]	1.32 [2]	3.32 [3]	3.51 [4]	3.98 [5]	4.73 [6]	6.15 [7]	8.02 [9]	6.76 [8]	*13.09* [10]

Benchmark criteria (×100): 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; BCE = bidirectional consistency error; GBCE = global bidirectional consistency error; BGM = bipartite graph matching; SC = segmentation covering; SSC = segmentation covering 2; VD = Van Dongen metric; L = Larsen metric; AVI = adjusted variation of information; NVI = normalized variation of information; NMI = normalized mutual information; M = Mirkin metric; ARI = adjusted Rand index; JC = Jaccard coefficient; DC = Dice coefficient; FMI = Fowlkes-Mallow index; WI = Wallace discrepancy I; WII = Wallace discrepancy II; NBDE = normalized boundary displacement error; superscripts = ranks; RANK, AVG, NORM = meta-criteria; values in bold are the best while italic values are the worst.

was computed from 180 segmentation results on the Prague texture benchmark color mosaics and 180 segmentations on the Berkeley data set. The Spearman rank correlation agrees well with the Berkeley data set and the Prague texture benchmark and confirms the Pearson correlation results in the previous section. The least rank correlated criteria on the both combined test sets are successive OS, US, O, NE, ME, $NBDE$, ... The OS criterion is, for example, uncorrelated with the following criteria:

$$CA, CO, I, MS, BCE, GBCE, BGM, SC, SSC, VD,$$
$$AVI, NVI, NMI, ARI, JC, DC, FMI, NBDE$$

on both sets and uncorrelated on separate sets on several other criteria.

4 Ranking Stability

Image segmentation papers, with a few honorable exceptions, present their methods on insufficient number of test images. Therefore, their results have a negligible information value. Fig. 3 – left (dotted lines) illustrates this problem on seven well-known segmenters[1] Blobword [5], DBM [30], EDISON [9], HGS (W) [34], SEG [13], EGBIS [20], TBES [54]) using the over-segmentation (OS) criterion performance on nine different test mosaic sets, each set having 20 mosaics. None of these segmenters has stable performance on all of these test sets. The OS criterion performance changes from 8 percent difference for the HGS (W) method [34] to 24 percent difference for JSEG [13] between different sets of 20 mosaics with the mean difference over six methods being 15 percent and standard

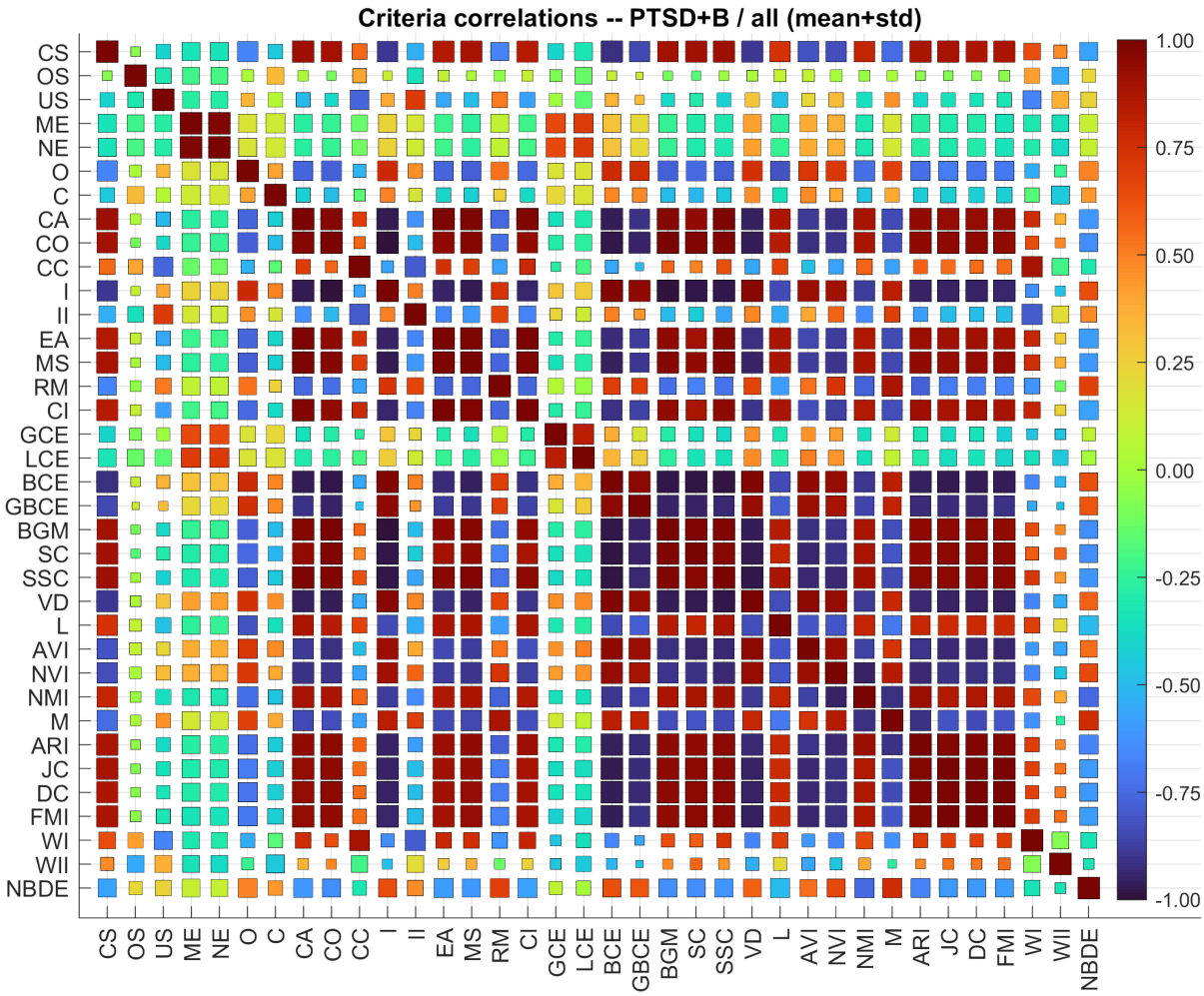

Fig. 2. Correlation between 36 segmentation criteria (mean correlation over 17 methods and 180 segmentation results each). A rectangle's color and size correspond to the correlation value and standard deviation, respectively.

deviation 5 percent. The correct detection (CS) changes from 7 percent difference for the EDISON method [9] to 18 percent difference for HGS (W) with the mean difference over six methods being 13 percent and the standard deviation 3 percent. The CS, OS, ME, NE, O, C, and MS criteria vary the most. The HGS (W) method has the largest variations for ten criteria, while TBES [54] has thirteen smallest criteria variations. The consequence is an unstable quality ranking between DBM [30] and Blobword [5] or JSEG and EDISON. Fig. 3 – right confirms this typical behavior – rank swapping using the over-segmentation (OS) criterion. Fig. 3 (solid lines) presents a typical segmenter behavior; the single quality criteria and the corresponding algorithm's ranking become stable only after at least eighty test images. Thus any segmenter validation is faithful only for large test sets. Small test sets can just suggest a possible behavior and its approximate ranking.

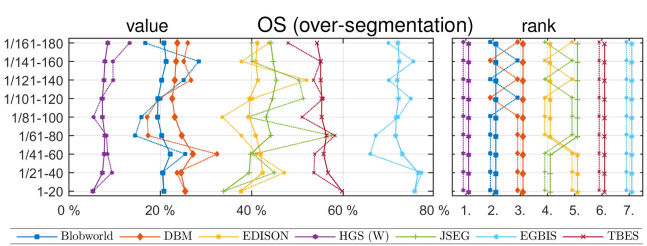

Fig. 3. Stability (dotted line averaged 20 test mosaics, solid line increasing test size from 20 to 180 mosaics) graphs for seven segmenters and the over-segmentation criterion.

5 Mosaics versus Real Images

Another natural question is how realistic such extensive performance evaluation is on computer assembled textural mosaics. Visual scenes contain objects from various materials; these materials are typically represented as visual textures [28], [40] mapped on the corresponding object shapes. Thus any real image can be represented as a textural mosaic. However, a material's appearance predominantly depends on the viewing, illumination, and shape properties, among other [28]. The viewing and illumination conditions are somewhat varied for each texture in the test mosaic, the viewing direction follows the surface normal, and all textures have correct natural illumination. This illumination is mostly solar and only approximately consistent between different textures of the mosaic. The test data are roughly planar and as such they only approximate a real visual scene with general object shapes with geometrically distorted material surface textures. However, they allow us to make use of the exact ideal and non-subjective segmentation, and to generate test sets of any size we wish. But, most importantly, the ranking of the

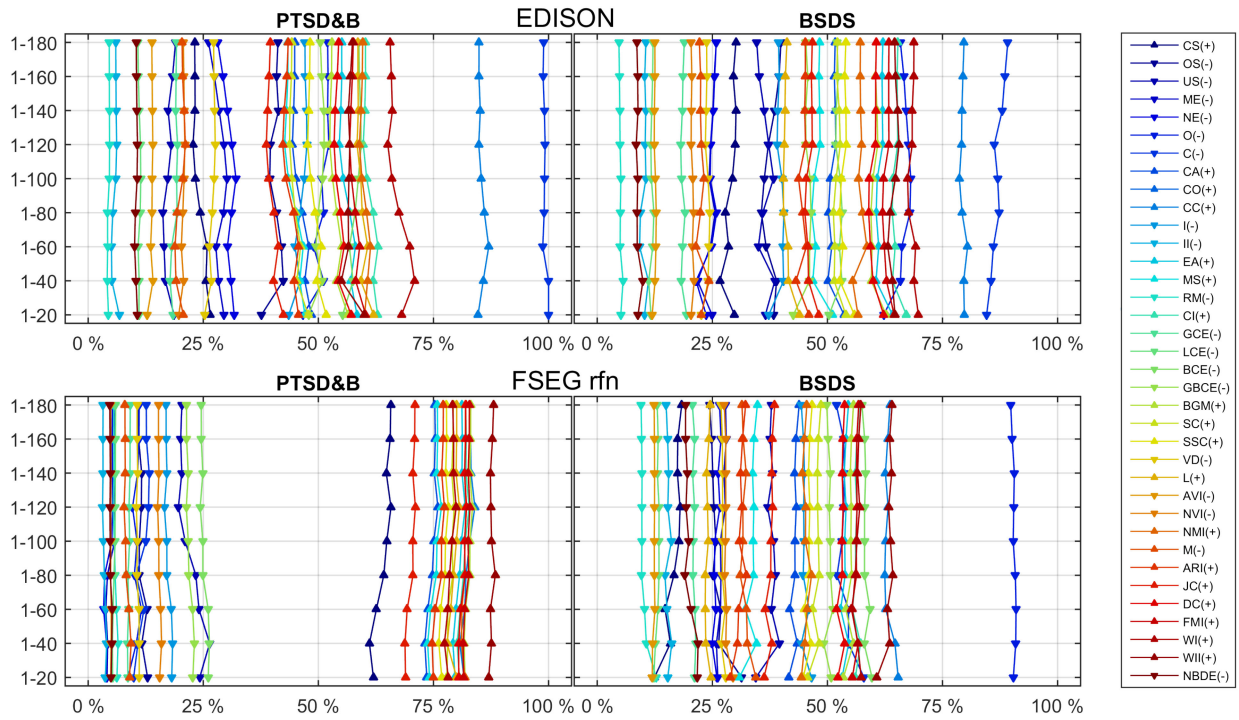

Fig. 4. Criteria comparison for EDISON [9] and FSEG [80] methods on benchmark contest (PTSD&B) [30] and Berkeley data sets (BSDS) for increasing test size from 20 to 180 images or mosaics. Arrow direction suggests the required criterion direction.

segmentation methods correlate well with the experiments on real natural scenes, as we have verified on the Berkeley test database [49], and the unlimited size of the test is crucial to obtain stable performance ranking. The Berkeley test database has up to five different subjective ground truth for each image; we thus compare the average correlation over all these alternatives with our mosaic results for numerous benchmark segmenters. The average Pearson's correlation between the Berkeley and our benchmark test sets is 0.87 (for seventeen unsupervised segmenters, 36 test criteria, and 180 test images in both databases). Similarly, the Spearman rank correlation agrees in 83 percent between the Berkeley data set and the Prague texture benchmark (for significance level 0.01, 36 test criteria, and 180 test images in both databases). Fig. 4 – top illustrates[1] this similarity for the EDISON method. Most of the 36 criteria curves have similar appearances on the presented benchmark data and on the Berkeley data set. The FSEG curves (Fig. 4 – bottom) significantly differ due to each method's tuning to the contest data [30]. All these experiments suggest that our benchmark is a robust image segmentation algorithm's evaluation tool. It simultaneously outperforms alternative benchmarks with its major superiority features - an unlimited number of user-controlled (e.g., number of regions, border shapes, noise content) test images and the objective ground truth. Such crucial properties cannot be achieved with any real image test sets such as BSDS500 [3] or Outex Texture Database [56].

6 Examples

Detailed analysis of single methods is beyond the scope of this article due to the nearly 200 different segmentation methods and several thousand test results recently present in the benchmark. The benchmark's functionalities are demonstrated in the evaluation of ten previously published examples of unsupervised or supervised image segmentation methods. However, they are used to illustrate the benchmark functionality and not to review the recent image segmentation state-of-the-art. The majority of segmentation methods are tested in the benchmark by their authors, which guarantees their evaluation objectivity, and every author can freely decide whether to keep his or her method's results in the benchmark database. We witnessed various methods of withdrawal from our benchmark often when the corresponding method's performance was only average and thus potentially endangering its possible publication. The following examples detail color textures/images because the vast majority of image segmenters aim for these data. Nevertheless, analogical analysis can be easily done for other types of data, only briefly discussed in the following subsections. All the corresponding details, both numerical and visual, in this section can be checked on the benchmark web and some comments on Table 1 results can be checked in [30]. We only discuss specific results for further study.

6.1 BTF Textures

BTF mosaics are created by BTF wood species measurements mapped on artificially created 3D surfaces. Each surface triangle is mapped with a physically correct wood material measurement which precisely corresponds to the local illumination and viewing conditions, and as such it represents the state-of-the-art realistic material visual representation [28] and eliminates most of the benchmark approximations (Section 5) in comparison with real visual scenes. It is simultaneously the only existing BTF segmentation benchmark. The

two best methods out of the nine tested on this benchmark are VRA-PMCFA and MW3AR8 [31], [58].

6.2 Hyperspectral Remote Sensing Textures

The remote sensing benchmark [53] uses the ten spectral bands (0.048–2.35 $[\mu m]$) Advanced Land Imager (ALI) [17] and the high-resolution (up to 0.41 m, three spectral bands) GeoEye observations [25]. The benchmark uses 31 multispectral ALI and 52 GeoEye color textures categorized into twelve thematic classes. These satellite mosaics are verified on numerous unsupervised, supervised, and several commercial remote sensing classifiers (for details check [53] or mosaic.utia.cas.cz/?act = view_res&bid = 18). A detailed evaluation shows the importance of the state-of-the-art textural features for good classifier performance or the Gaussian-mixture-model-based clustering over more straightforward standard clustering metrics. High-resolution data are particularly challenging, and all the benchmarked techniques perform uniformly worse than on the ALI dataset.

6.3 Dynamic Textures

Color dynamic textures have two types of dynamics – variable regions both in shape and location and single dynamic regional textures from the DynTex database [62]. This benchmark recently contains a comparison of 4 methods [29] and six modifications of one of them. A detailed evaluation is available in the benchmark (mosaic.utia.cas.cz/?act = view_res&dyn = 1).

6.4 Color Textures

The benchmark performance is demonstrated by comparing six unsupervised and four supervised (where *, † denote either using supplemented training data or manually added learning information) segmentation algorithms in Table 1 - these ten recently published methods are CMS [59], †RS [81], IGMRF [14], LGG [79], PCA-MS [52], GRPNMF [4], two *,†FCNT variants [2], A3M [43], and *EWT-FCNT [37]. The performance details of numerous other methods and further details (performance criteria, curves, all test mosaics segmentations, etc.) can be found on the benchmark server.

Almost every algorithm has several parameters to be tuned (e.g., weights, thresholds, seeds, and some other) and they often significantly influence its segmentation performance and the corresponding quality criteria. All segmentation results stored in the benchmark were produced either with a default parameter setting suggested by their authors or with the best parameters set tuned to the benchmark data. However, the benchmark criteria are computed from at least 20 experimental segmentations, so it is not easy to artificially tune parameters to produce atypically outstanding results and thus a biased ranking of a preferred method.

The unsupervised k-means clustering method [14] IGMRF uses a texture descriptor named local parameter histograms. These features are computed from Gaussian Markov random fields (GMRF) local estimates. IGMRF addresses the inconsistencies arising in localized parameter estimation by applying generalized inverse, regularization, and an estimation window size selection criterion.

The supervised segmentation method †RS [81] uses the least square solution of the linear regression model. †RS uses local spectral histograms as the feature vectors. These vectors consist of histograms of the intensity filter, two LoG (Laplacian of Gaussian) filters with the scale values of 0.2 and 0.5, and four Gabor filter responses. The filterbanks and integration scales are given manually. The method is limited to small grayscale images due to the linear regression over the whole image and all features.

An unsupervised LGG method [79] is a global/local affinity sparse graph-cut image segmentation over superpixels. Global grouping is achieved using medium-sized superpixels through a sparse representation of superpixel's features. Small- and large-sized superpixels are then used to achieve local smoothness through an adjacent graph in a given color histogram, LBP, and SIFT feature space. Different graphs are heuristically fused. The method is prolonged. It needs 31 minutes to segment a single Berkely image.

An A3M segmenter [43] is a framework to learn convolutional features based on the piecewise constant Mumford-Shah model for unsupervised texture segmentation when no training data is available. The underlying idea is to learn suitable filters in a way such that their responses (after applying the non-linearity) on the segments is approximately constant.

*,†FCNT [2] is a supervised segmentation method which uses a fully convolutional FCN8 network with four convolutional layers. Moreover, it combines the response to filter banks at various depths. The region boundaries are localized by combining local and global information in the deconvolution layers. †FCNT uses the A3M segmenter to obtain learning data.

The supervised segmentation method *EWT-FCNT [37] uses a fully convolutional network. The texture features are extracted from images using an empirical curvelet transform. Each image has its own set of curvelet filters. These features are subsequently fed into a fully convolutional network.

The unsupervised PCA-MS [52] segmentation utilizes the multi-phase Mumford-Shah model. The high-dimensional textural features in the form of local spectral histograms of Gabor features are projected onto a low-dimensional space using the principle component analysis.

An unsupervised graph clustering and image segmentation algorithm GRPNMF [4] uses the projective nonnegative matrix factorization to lessen the representation dimensionality of the Lab color and Gabor filter bank on five scales and eight orientation features.

The unsupervised method CMS is based on the cooperative region merging [59].

Fig. 6 – the first column and odd rows show three selected 512×512 mosaics from the color benchmark created from three to ten natural color textures and the even rows contain their corresponding ground truth, i.e., the ideal segmentation. The last five subsequent columns in Fig. 6 demonstrate comparative results from ten alternative algorithms – *EWT-FCNT, *FCNT, †FCNT, A3M, PCA-MS, GRPNMF, CMS, LGG, IGMRF, and †RS. The visual comparison suggests the inclination for IGMRF over-segmentation and, to less extent, also in †RS, and largely missed and noise errors in LGG.

LGG (Table 1) indicates the worst both missed and noise errors. The *EWT-FCNT method has the best average rank from all compared methods as well as the best performing criteria (all presented 36). A3M has the best average rank

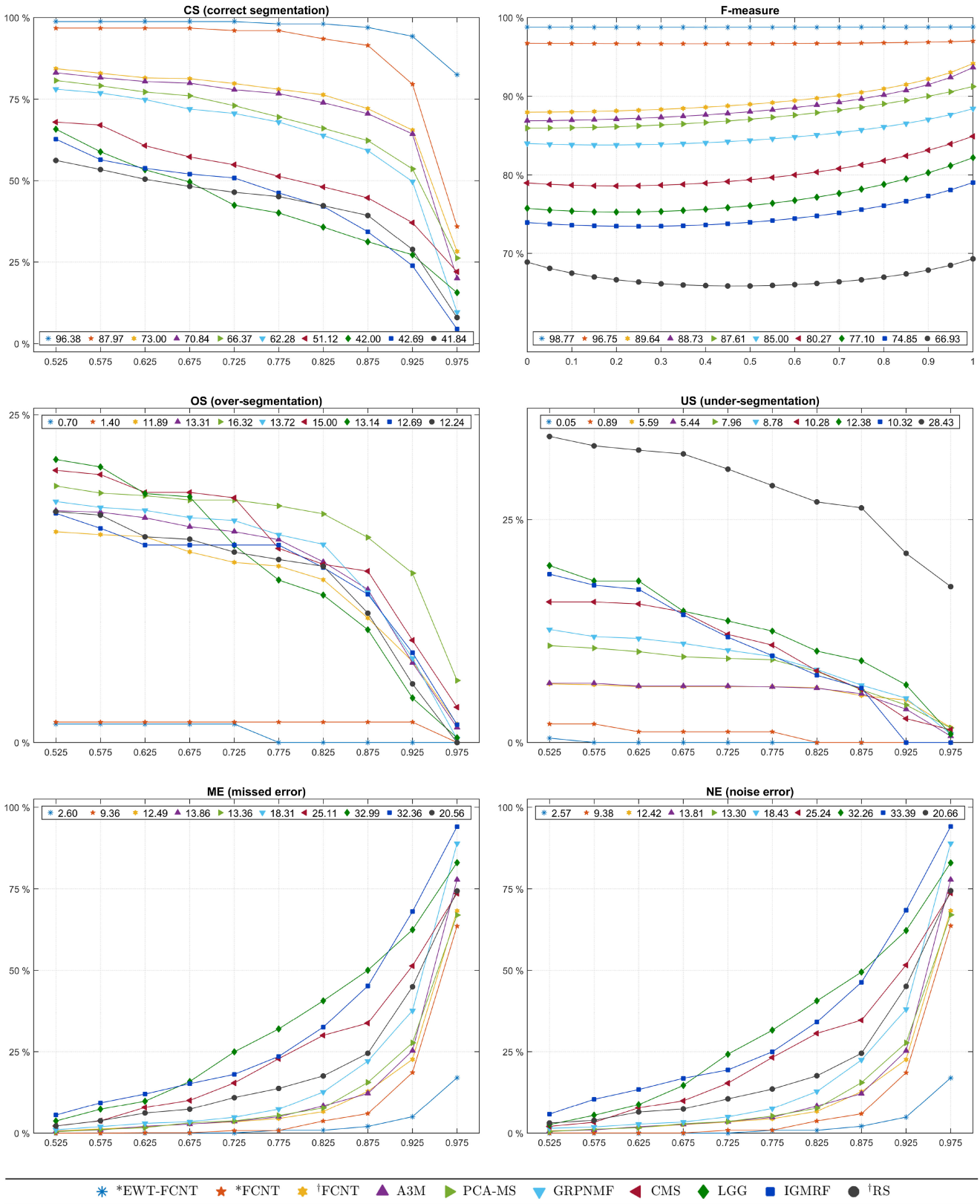

Fig. 5. Performance curves and the corresponding performance integrals for *EWT-FCNT, *FCNT, †FCNT, A3M, PCA-MS, GRPNMF, GRPNMF, Cooperative Mum-Shah (CMS), Local Global Graph Cut (LGG), Improved GMRF (IGMRF), †RS methods averaged over 20 or 80 mosaics and increasing threshold ($t \in \langle 0.525; 0.975 \rangle$).

from unsupervised methods and thirty performing criteria better than the remaining five unsupervised methods and even the worst performing the supervised †RS method, which has only the OS criterion better.

Fig. 6 – the second and fifth columns (odd rows) demonstrate robust behavior of the *EWT-FCNT supervised and A3M unsupervised segmenters but also infrequent A3M failures caused by producing over-segmented thematic maps for

Fig. 6. Three selected texture mosaics from the benchmark with the corresponding ground-truth, and segmentation results for *EWT-FCNT, *FCNT, †FCNT, A3M, PCA-MS, GRPNMF, Cooperative Mum-Shah (CMS), Local Global Graph Cut (LGG), Improved GMRF (IGMRF), †RS, respectively.

some textures. A more elaborate post-processing step can correct such failures. The PCA-MS, GRPNMF, CMS, LGG, IGMRF, and †RS algorithms performed steadily worse on the benchmark data, which can also be checked in Table 1.

The integrated numerical results over the whole color benchmark (20/80 different mosaics) in Fig. 5 confirm these observations. *EWT-FCNT (A3M unsupervised) produces the best correct segmentation, followed by *FCNT and †FCNT, while †RS is the worst. PCA-MS, CMS, and GRPNMF (Table 1) have strong over-segmentation (OS) tendency though lower ME, NE errors for PCA-MS. A3M confirms the best inter-region border localization of these unsupervised methods. However, A3M has slightly less precisely located borders than the best *EWT-FCNT supervised method. †RS has strong under-segmentation (US) performance.

The performance curves show the sensitivity to threshold parameter for the CMS, †RS, IGMRF, LGG, PCA-MS, GRPNMF, two *,†FCNT variants, A3M, and *EWT-FCNT methods in Fig. 5. Similarly, their integrals (Fig. 5 in all graphs for all methods) confirm for the best methods from Table 1 that their behavior is not too sensitive concerning the region-based criteria threshold. Fig. 5 (CS) simultaneously shows the most precise regions border localization of the *EWT-FCNT

supervised (A3M unsupervised) method, similar precision of the PCA-MS method, and finally, the worst localization of the IGMRF, †RS methods. As expected, this threshold mainly affects the inter-region border localization. The localization error difference between the best and the worst method has been only slightly diminished over the whole threshold range. The pixel-wise criteria (omission error, recall, etc.) further confirm the superiority of both *EWT-FCNT and A3M methods in their categories. PCA-MS leads over A3M with a small ratio of noise error (NE) and in the O, C, AVI, and LCE criteria.

The overall conclusion supports the superiority of the *EWT-FCNT and A3M methods over the remaing tested methods in their corresponding categories. The algorithms evaluation can be further supplemented with other important attributes such as the noise sensitivity, processing speed and difficulties by their corresponding parameter setting.

7 Conclusion

The implemented supervised/unsupervised segmentation benchmark is a fully automatic web application, which enables us to compare mutually image segmentation algorithms and to assist in developing new segmentation methods. The comparison can be made for finalized algorithms with results, descriptions, and references stored permanently in the benchmark database and used for subsequent comparison with other algorithms or for a working version of a segmenter. Segmenters can be ranked based on a chosen criterion from the set of over forty region, pixel, consistency, set, information, border, or clustering based criteria or any subset with user stated criteria relative importance. The test mosaics, as well as the ground truths, are computer-generated, which guarantees the evaluation objectivity and allows for easy generation of extensive test sets which are otherwise unfeasible to arrange. The benchmark enables us to test single algorithms on monospectral, multispectral, BTF or dynamic 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, objectively, and effectively compare their novel algorithms and verify their performance characteristics.

Among important aspects which are not currently tested is mainly the resilience against complex geometric distortions (e.g., foreshortening) and segmentation speed, which cannot be tested because the benchmark only analyzes the uploaded segmentation results, which were achieved on a wide range of varied computer architectures by various developers worldwide. Only a subset of the method's results was submitted together with their original code. The most segmentation methods in the benchmark are inserted by their authors, which guarantees their evaluation objectivity and every author can freely decide whether to keep his or her method's results in the benchmark database or not. Thus some methods may even disappear either due to their inferior performance or because their authors prefer to release these results in some later publication.

Although the benchmark is primarily designed for texture segmenters, it gives helpful performance insight for any general image segmenter. The evaluation part of the benchmark is also modified to utilize user-defined ground truth, such as hand segmented natural images. However, such results will not be stored in the benchmark database, and hence they will not be available for comparison to other users. Other possible applications, such as machine learning methods evaluation, the wrapper of filter-based feature selection method comparison, image compression testing, query by pictorial example method evaluation and some others can easily benefit from the benchmark services as well.

The usefulness of the benchmark is acknowledged in over a hundred publications using the benchmark results.

Acknowledgments

This work was supported by the Czech Science Foundation Project GAČR under Grant 19-12340S .

References

[1] S. Alpert, M. Galun, A. Brandt, and R. Basri, "Image segmentation by probabilistic bottom-up aggregation and cue integration," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 2, pp. 315–327, Feb. 2012.

[2] V. Andrearczyk and P. F. Whelan, "Texture segmentation with fully convolutional networks," 2017, *arXiv:1703.05230*.

[3] P. Arbelàez, M. Maire, C. Fowlkes, and J. Malik, "Contour detection and hierarchical image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 5, pp. 898–916, May 2011.

[4] C. G. Bampis, P. Maragos, and A. C. Bovik, "Projective non-negative matrix factorization for unsupervised graph clustering," in *Proc. IEEE Int. Conf. Image Process.*, 2016, pp. 1255–1258.

[5] 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 *Proc. Int. Conf. Comput. Vis.*, 1998, pp. 675–682.

[6] M. Borsotti, P. Campadelli, and R. Schettini, "Quantitative evaluation of color image segmentation results," *Pattern Recognit. Lett.*, vol. 19, no. 8, pp. 741–747, Jun. 1998.

[7] S. Chabrier, B. Emile, C. Rosenberger, and H. Laurent, "Unsupervised performance evaluation of image segmentation," *EURASIP J. Appl. Signal Process.*, vol. 2006, 2006, Art. no. 096306.

[8] H. Cheng, X. Jiang, Y. Sun, and J. Wang, "Color image segmentation: Advances and prospects," *Pattern Recognit.*, vol. 34, no. 12, pp. 2259–2281, 2001.

[9] C. Christoudias, B. Georgescu, and P. Meer, "Synergism in low level vision," in *Proc. 16th Int. Conf. Pattern Recognit.*, vol. 4, 2002, pp. 150–155.

[10] M. Cordts *et al.*, "The cityscapes dataset for semantic urban scene understanding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 3213–3223.

[11] P. Correia and F. Pereira, "Objective evaluation of relative segmentation quality," in *Proc. Int. Conf. Image Process.*, vol. 1, 2000, pp. 308–311.

[12] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 248–255.

[13] Y. Deng and B. Manjunath, "Unsupervised segmentation of color-texture regions in images and video," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 8, pp. 800–810, Aug. 2001.

[14] C. Dharmagunawardhana, S. Mahmoodi, M. Bennett, and M. Niranjan, "Gaussian Markov random field based improved texture descriptor for image segmentation," *Image Vis. Comput.*, vol. 32, no. 11, pp. 884–895, 2014.

[15] L. R. Dice, "Measures of the amount of ecologic association between species," *Ecology*, vol. 26, no. 3, pp. 297–302, 1945.

[16] A. Dutta and A. Zisserman, "The VIA annotation software for images, audio, and video," in *Proc. 27th ACM Int. Conf. Multimedia*, 2019, pp. 2276–2279.

[17] Earth Resources Observation And Science (EROS) Center, "Earth Observing One (EO-1) - ALI," 2000. [Online]. Available: https://www.usgs.gov/centers/eros/science/usgs-eros-archive-earth-observing-one-eo-1-ali

[18] M. Everingham, H. Muller, and B. Thomas, "Evaluating image segmentation algorithms using the Pareto front," in *Proc. 7th Eur. Conf. Comput. Vis.-Part IV*, 2002, pp. 34–48.

[19] M. Everingham, L. Van Gool , C. K. Williams, J. Winn, and A. Zisserman, "The PASCAL visual object classes (VOC) challenge," *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 303–338, 2010.

[20] P. Felzenszwalb and D. Huttenlocher, "Efficient graph-based image segmentation," *Int. J. Comput. Vis.*, vol. 59, no. 2, pp. 167–181, Sep. 2004.

[21] E. B. Fowlkes and C. L. Mallows, "A method for comparing two hierarchical clusterings," *J. Amer. Stat. Assoc.*, vol. 78, no. 383, pp. 553–569, 1983.

[22] A. L. N. Fred and A. K. Jain, "Robust data clustering," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, , 2003, vol 2, pp. II–II.

[23] J. Freixenet, X. Munoz, D. Raba, J. Marti, and X. Cufi, "Yet another survey on image segmentation: Region and boundary information integration," in *Proc. Eur. Conf. Comput. Vis.*, 2002, pp. 408–422.

[24] M. Galun, E. Sharon, R. Basri, and A. Brandt, "Texture segmentation by multiscale aggregation of filter responses and shape elements," in *Proc. Int. Conf. Comput. Vis.*, 2003, pp. 716–723.

[25] GeoEye, Inc., "Geoeye product guide v1.0.1," 2009. [Online]. Available: http://www-igm.univ-mlv.fr/ riazano/enseignement/SR-FUSION-COURS/GeoEye_Product_Guide.pdf

[26] M. Haindl, J. Filip, R. Vávra, and S. Mikeš, "UTIA bidirectional texture function database," 2009. [Online]. Available: http://btf.utia.cas.cz

[27] M. Haindl, J. Grim, and S. Mikes, "Texture defect detection," in *Proc. Comput. Anal. Images Patterns*, 2007, pp. 987–994.

[28] M. Haindl and J. Filip, *Visual Texture (Advances in Computer Vision and Pattern Recognition)*. London, U.K.: Springer, Jan. 2013.

[29] M. Haindl and S. Mikeš, "Unsupervised dynamic textures segmentation," in *Computer Analysis of Images and Patterns (Lecture Notes in Computer Science 8047)*, R. Wilson, E. Hancock, A. Bors, and W. Smith, Eds., Berlin, Germany: Springer, Aug. 2013, pp. 433–440.

[30] M. Haindl and S. Mikeš, "A competition in unsupervised color image segmentation," *Pattern Recognit.*, vol. 57, no. 9, pp. 136–151, Sep. 2016.

[31] M. Haindl, S. Mikeš, and M. Kudo, "Unsupervised surface reflectance field multi-segmenter," in *Computer Analysis of Images and Patterns (Lecture Notes in Computer Science 9256)*, G. Azzopardi and N. Petkov, Eds., Berlin, Germany: Springer Int. Publishing, Sep. 2015, pp. 261–273.

[32] M. Haindl, S. Mikeš, and G. Scarpa, "Unsupervised detection of mammogram regions of interest," in *Knowledge-Based Intelligent Information and Engineering Systems (LNAI 4694)*, B. Apolloni, R. J. Howlett, and L. Jain, Eds., Berlin, Germany: Springer, 2007, pp. 33–40.

[33] R. M. Haralick and L. G. Shapiro, "Image segmentation techniques," *Comput. Vis., Graph. Image Process.*, vol. 29, no. 1, pp. 100–132, 1985.

[34] M. A. Hoang, J.-M. Geusebroek, and A. W. Smeulders, "Color texture measurement and segmentation," *Signal Process.*, vol. 85, no. 2, pp. 265–275, 2005.

[35] 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.

[36] Q. Huang and B. Dom, "Quantitative methods of evaluating image segmentation," in *IEEE Proc., Int. Conf. Image Process.*, 1995, vol. 3, pp. 53–56.

[37] Y. Huang, F. Zhou, and J. Gilles, "Empirical curvelet based fully convolutional network for supervised texture image segmentation," *Neurocomputing*, vol. 349, pp. 31–43, 2019.

[38] L. Hubert and P. Arabie, "Comparing partitions," *J. Classification*, vol. 2, no. 1, pp. 193–218, 1985.

[39] D. P. Huttenlocher, G. A. Klanderman, and W. A. Rucklidge, "Comparing images using the Hausdorff distance," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 15, no. 9, pp. 850–863, Sep. 1993.

[40] D. E. Ilea and P. F. Whelan, "Image segmentation based on the integration of colour-texture descriptors–A review," *Pattern Recognit.*, vol. 44, no. 10/11, pp. 2479–2501, 2011.

[41] P. Jaccard, "The distribution of the flora in the alpine zone," *New Phytologist*, vol. 11, no. 2, pp. 37–50, 1912.

[42] X. Jiang, C. Marti, C. Irniger, and H. Bunke, "Distance measures for image segmentation evaluation," *EURASIP J. Appl. Signal Process.*, vol. 2006, 2006, Art. no. 035909.

[43] M. Kiechle, M. Storath, A. Weinmann, and M. Kleinsteuber, "Model-based learning of local image features for unsupervised texture segmentation," *IEEE Trans. Image Process.*, vol. 27, no. 4, pp. 1994–2007, Apr. 2018.

[44] B. Larsen and C. Aone, "Fast and effective text mining using linear-time document clustering," in *Proc. 5th ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 1999, pp. 16–22.

[45] S. U. Lee, S. Y. Chung, and R. H. Park, "A comparitive performance study of several global thresholding techniques for segmentation," *Comput. Vis., Graph. Image Process.*, vol. 52, no. 2, pp. 171–190, 1990.

[46] M. D. Levine and A. M. Nazif, "Dynamic measurement of computer generated image segmentations," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-7, no. 2, pp. 155–164, Mar. 1985.

[47] T.-Y. Lin *et al.*, "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 740–755.

[48] L. Lucchese and S. Mitra, "Color image segmentation: A state-of-the-art survey," *Proc. Indian Nat. Sci. Acad.*, vol. 67, no. 2, pp. 207–221, 2001.

[49] 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.*, Jul. 2001, vol 2, pp. 416–423. [Online]. Available: http://www.cs.berkeley.edu/projects/vision/grouping/segbench/

[50] D. R. Martin, J. Malik, and D. Patterson, "An empirical approach to grouping and segmentation," Ph.D. dissertation, Comput. Sci. Dept., Univ. California, Berkeley, USA, 2002.

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

[52] N. Mevenkamp and B. Berkels, "Variational multi-phase segmentation using high-dimensional local features," in *Proc. IEEE Winter Conf. Appli. Comput. Vis.*, 2016, pp. 1–9.

[53] S. Mikes, M. Haindl, G. Scarpa, and R. Gaetano, "Benchmarking of remote sensing segmentation methods," *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.*, vol. 8, no. 5, pp. 2240–2248, May 2015.

[54] H. Mobahi, S. R. Rao, A. Y. Yang, S. S. Sastry, and Y. Ma, "Segmentation of natural images by texture and boundary compression," *Int. J. Comput. Vis.*, vol. 95, no. 1, pp. 86–98, 2011.

[55] J. Munkres, "Algorithms for the assignment and transportation problems," *J. Soc. Ind. Appl. Math.*, vol. 5, no. 1, pp. 32–38, Mar. 1957.

[56] T. Ojala, T. Maenpaa, M. Pietikainen, J. Viertola, J. Kyllonen, and S. Huovinen, "Outex: New framework for empirical evaluation of texture analysis algorithms," in *Proc. Int. Conf. Pattern Recognit.*, 2002, pp. 701–706.

[57] N. Pal and S. Pal, "A review on image segmentation techniques," *Pattern Recognit.*, vol. 26, no. 9, pp. 1277–1294, 1993.

[58] C. Panagiotakis, I. Grinias, and G. Tziritas, "Natural image segmentation based on tree equipartition, Bayesian flooding and region merging," *IEEE Trans. Image Process.*, vol. 20, no. 8, pp. 2276–2287, Aug. 2011.

[59] F. C. Patino, "Information theoretical region merging approaches and fusion of hierarchical image segmentation results," Ph.D. dissertation, Signal Theory Commun. Dept., Universitat Politecnica de Catalunya, Barcelona, Spain, Feb. 2010.

[60] N. Payet and S. Todorovic, "Hough forest random field for object recognition and segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 5, pp. 1066–1079, May 2013.

[61] F. Perazzi, J. Pont-Tuset , B. McWilliams, L. Van Gool , M. Gross, and A. Sorkine-Hornung , "A benchmark dataset and evaluation methodology for video object segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 724–732.

[62] R. Peteri, S. Fazekas, and M. J. Huiskes, "DynTex: A comprehensive database of dynamic textures," *Pattern Recognit. Lett.*, vol. 31, no. 12, pp. 1627–1632, 2010.

[63] J. Pont-Tuset and F. Marques, "Supervised evaluation of image segmentation and object proposal techniques," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 7, pp. 1465–1478, Jul. 2016.

[64] W. M. Rand, "Objective criteria for the evaluation of clustering methods," *J. Amer. Stat. Assoc.*, vol. 66, no. 336, pp. 846–850, 1971.

[65] P. K. Sahoo, S. Soltani, and A. K. C. Wong, "Survey of thresholding techniques," *Comput. Vis., Graphi. Image Process.*, vol. 41, no. 2, pp. 233–260, 1988.

[66] M. Sattler, R. Sarlette, and R. Klein, "Efficient and realistic visualization of cloth," in *Proc. Eurographics Symp. Rendering* , Jun. 2003, pp. 167–177.

[67] G. Scarpa, M. Haindl, and J. Zerubia, "A hierarchical texture model for unsupervised segmentation of remotely sensed images," *Lecture Notes Comput. Sci.*, vol. 4522, pp. 303–312, 2007.

[68] C. W. Shaffrey, I. H. Jermyn, and N. G. Kingsbury, "Psychovisual evaluation of image segmentation algorithms," in *Proc. Adv. Concepts Intell. Vis. Syst.*, 2002.

[69] M. Sharma and S. Singh, "Minerva scene analysis benchmark," in *Proc. 7th Australian New Zealand Intell. Inf. Syst. Conf.*, 2001, pp. 231–235.
[70] E. Shelhamer, J. Long, and T. Darrell, "Fully convolutional networks for semantic segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 640–651, Apr. 2017.
[71] J. R. Shewchuk, "Triangle: Engineering a 2D quality mesh generator and delaunay triangulator," in *Applied Computational Geometry: Towards Geometric Engineering (Lecture Notes in Computer Science 1148)*, M. C. Lin and D. Manocha, Eds. Berlin, Germany: Springer, May 1996, pp. 203–222.
[72] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 8, pp. 888–905, Aug. 2000.
[73] S. Srubar, "Speed comparison of segmentation evaluation methods," in *Combinatorial Image Analysis*, R. P. Barneva, V. E. Brimkov, and J. Šlapal, Eds. Cham, Switzerland: Springer, 2014, pp. 113–122.
[74] A. Strehl and J. Ghosh, "Cluster ensembles — A knowledge reuse framework for combining multiple partitions," *J. Mach. Learn. Res.*, vol. 3, no. 3, pp. 583–617, 2003.
[75] R. Unnikrishnan, C. Pantofaru, and M. Hebert, "A measure for objective evaluation of image segmentation algorithms," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2005, pp. 34–34.
[76] S. Wagner and D. Wagner, "Comparing clusterings – An overview," Universität Karlsruhe, Karlsruhe, Germany, *Tech. Rep. 2006–04*, 2007.
[77] D. L. Wallace, "A method for comparing two hierarchical clusterings: Comment," *J. Amer. Stat. Assoc.*, vol. 78, no. 383, pp. 569–576, 1983.
[78] M. Wang, X. Liu, Y. Gao, X. Ma, and N. Q. Soomro, "Superpixel segmentation: A benchmark," *Signal Process., Image Commun.*, vol. 56, pp. 28–39, 2017.
[79] X. Wang, Y. Tang, S. Masnou, and L. Chen, "A global/local affinity graph for image segmentation," *IEEE Trans. Image Process.*, vol. 24, no. 4, pp. 1399–1411, Apr. 2015.
[80] J. Yuan and D. Wang, "Factorization-based texture segmentation," Ohio State Univ., Columbus, USA, *Tech. Rep. OSU-CISRC-1/13-TR0*, 2013.
[81] J. Yuan, D. Wang, and R. Li, "Image segmentation based on local spectral histograms and linear regression," in *Proc. Int. Joint Conf. Neural Netw.*, 2011, pp. 482–488.
[82] H. Zhang, J. E. Fritts, and S. A. Goldman, "Image segmentation evaluation: A survey of unsupervised methods," *Comput. Vis. Image Understanding*, vol. 110, no. 2, pp. 260–280, 2008.
[83] H. Zhang, J. E. Fritts, and S. A. Goldman, "An entropy-based objective evaluation method for image segmentation," in *Proc. SPIE- Storage Retrieval Methods Appli. Multimedia*, 2004, vol. 5307, pp. 38–49.
[84] Y. J. Zhang, "A review of recent evaluation methods for image segmentation," in *Signal Process. Appl.*, 2001, vol. 1, pp. 148–151.
[85] Y. Zhang, "A survey on evaluation methods for image segmentation," *Pattern Recognit.*, vol. 29, no. 8, pp. 1335–1346, Aug. 1996.
[86] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba, "Scene parsing through ADE20K dataset," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 5122–5130.

Stanislav Mikeš received the Graduation degree and the PhD degree from the Faculty of Mathematics and Physics, Charles University, Prague, in 2002 and 2010, respectively. His research interests include visual texture modeling, segmentation, benchmarking, and pattern recognition.

Michal Haindl (Senior Member, IEEE) received the graduation degree from Czech Technical University, Prague in 1979, the PhD degree from the Czechoslovak Academy of Sciences in 1983, and the ScD degree in 2001. Since 1983, he has been working on various image analysis and pattern recognition topics with the Institute of Information Theory and Automation (UTIA), Czechoslovak Academy of Sciences, Prague, University of Newcastle, Rutherford Appleton Laboratory, CWI, Amsterdam, and the INRIA, Rocquencourt. In 1995, he rejoined UTIA, where he is currently the head of Pattern Recognition Department. His current research interests include random fields applications in pattern recognition and image processing. He is an IAPR fellow and a professor.