Systematic Benchmarking of Aerial Image Segmentation

Jiangye Yuan, *Member, IEEE*, Shaun S. Gleason, *Member, IEEE*, and Anil M. Cheriyadat, *Member, IEEE*

***Abstract*—This letter presents a benchmarking study for aerial image segmentation. We construct an image data set consisting of various aerial scenes. Segmentations generated by different human subjects are used as ground truth. We analyze the consistency between segmentations from different subjects. We select six leading segmentation algorithms, which include not only the algorithms specifically designed for aerial images but also more generally applicable algorithms. We also select a recently proposed algorithm due to its promising performance in handling texture regions. We apply these algorithms to the aerial image data set and quantitatively evaluate their performance. We interpret the evaluation results based on the characteristics of algorithms, which provide general guidance for selecting proper algorithms in specific applications.**

***Index Terms*—Aerial image dataset, image segmentation.**

I. Introduction

IMAGE segmentation aims to partition an image into different regions such that each region should be as homogeneous as possible and neighboring ones should be as different as possible. Even though the segmentation objective is simple, the definition for a good segment can range from spatially contiguous pixels with uniform intensity, color, or texture to group of pixels representing "visually meaningful" objects. The diversity in the possible solutions makes it difficult to measure and optimize segmentation solutions. Often, aerial image segmentation algorithms are evaluated with respect to different measures such as classification accuracies and detection rates. To achieve more simple and generalized segmentation performance benchmarking, we systematically measure the performance with respect to human segmentations.

The Berkeley segmentation data set (BSD) [1], which consists of natural images, is often used to benchmark segmentation algorithms. However, due to very different data characteristics, the benchmarking results for natural images cannot apply to remote sensing images. There also exist some segmentation benchmarking studies in the remote sensing community. For instance, Meinel and Neubert [2] assess segmentation quality of seven segmentation programs using two multispectral images of size 2000 $\times$ 2000 pixels. Clinton *et al.* [3] investigate different measures for evaluating segmentations on a single urban image with two different segmentation software packages. However, these studies use a limited amount of benchmark data and overlook advanced methods developed for general image segmentation.

The main contribution of this letter is to objectively evaluate the performance of different segmentation algorithms on aerial images. We first present an aerial image data set along with human-generated ground truth. Seven segmentation algorithms are selected, which have demonstrated promising performance in segmenting general images or remote sensing images. The algorithms are applied to the data set, and the results are quantitatively assessed. We intend to provide readers with criteria of selecting the most suitable methods for their specific tasks. We also hope to give insight for further improvement on aerial image segmentation.

II. Benchmark Dataset and Quantitative Measures

We collected 80 high-resolution aerial images with spatial resolutions ranging from 0.3 to 1.0 meter. The image set contains different scenes, including school, residential, city, warehouse, and power plants. The size of each image is 512 $\times$ 512 pixels.

We use the segmentations generated by human subjects as ground truth. Four human subjects were assigned to manually segment the images. Two subjects have image analysis background, and the other two not. Subjects were told to segment the images intelligently. No particular cues were provided, because we attempted to obtain general segmentations without any prior biases on what features or objects are more important than others. Fig. 1 shows six images in our data set and the corresponding human segmentations that are superimposed on each other. As we expect, there are wide variations among different human segmentations of the same image. However, a considerable consistency can still be observed. The consistency was analyzed using quantitative measures.

Our benchmarking strategy involves comparing machine generated segmentations with human segmentations under different evaluation metrics. We have identified three metrics, which are widely adopted in segmentation literature. The precision-and-recall framework proposed in [4] is used to evaluate boundary localization. The framework computes an explicit correspondence of machine and human boundary pixels, which leads to the counts of hits and misses given a certain amount of tolerance. Precision indicates the portion of the true positive among all detected boundaries, and recall represents how many boundaries in the ground truth are detected. The F-measure, defined as the harmonic mean of precision and recall, provides a summary statistic. We also use two region-based metrics. Probabilistic Rand Index (PRI) [5] measures the probability that a pair of samples has consistent labels between

Manuscript received March 1, 2013; revised April 4, 2013; accepted April 26, 2013. This work was supported in part by U.S. Department of Energy/National Nuclear Security Administration under Grant DOE-NNSA/NA-22.

The authors are with Oak Ridge National Laboratory, Oak Ridge, TN 37831 USA (e-mail: yuanj@ornl.gov; gleasonss@ornl.gov; cheriyadatam@ornl.gov).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/LGRS.2013.2261453

Fig. 1. Examples of aerial image data set. The first row shows six example images. The second row overlays segment boundaries generated by four subjects, where darker pixels correspond to the boundaries marked by more subjects.

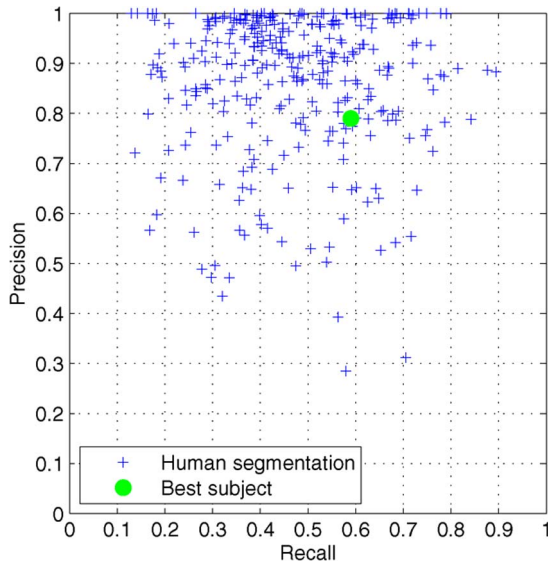

Fig. 2. Precision and recall of human segmentations. The points marked by a "+" show the precision and recall of all human segmentations for all images in the data set. The green dot represents the best average score among subjects.

two segmentations, and has large values if two segmentations are similar. Variation of Information (VOI) [6] measures the information difference between two segmentations, and gives small values if two segmentations are similar. In the case of multiple ground truth, the average score is used.

To quantitatively evaluate the consistency of human segmentations, for each image we compare the segmentation from one subject against those from the other three subjects and compute the precision and recall. Fig. 2 shows the scores of all segmentations from four subjects. The green dot shows the average precision and recall of segmentations from one subject, who achieves the highest F-measure among all subjects. We can see that the precision is much higher than the recall, which implies that, despite disagreement on the existence of certain boundaries, the detected boundaries by one subject are well aligned with those by other subjects.

III. Segmentation Algorithm Benchmarking

In our benchmarking study, we selected six segmentation algorithms that are representative of major families of segmentation approaches. They include JSEG [7], mean shift [8], the multi-resolution region merging algorithm (MSEG) [9], statistical region merging (SRM) [10], the graph-based region merging algorithm (Felz-Hutt) [11], and oriented watershed transform ultrametric contour maps with globalPb as contour detector (gPb-owt-ucm) [12]. We also include a recently proposed factorization-based segmentation algorithm [13]. These algorithms were chosen for the following reasons. First, these algorithms have shown promising performance on many image data set s. Secondly, they are among the most efficient segmentation algorithms. Computation time is an important concern when processing aerial images, which tend to be of a large size. The computational cost of gPb-owt-ucm is relatively high, which is selected because of its state-of-the-art performance on natural image segmentation. Finally, all these algorithms are capable of producing segmentations at different scale levels. This is highly desirable due to the fact that semantic regions in aerial images can have different characteristic scales. We use the codes of the algorithms distributed by the corresponding authors, if they are available. For mean shift, MSEG, and SRM, we use the implementations available at MeanShiftSrc [14], MSEGSrc [15], and SRMSrc [16], respectively. For each of the algorithms we vary certain scale parameters producing different segmentations. Each segmentation is quantitatively measured against the human segmentation set. In the following, we will discuss technical details of the algorithms.

The JSEG algorithm uses two steps to obtain segmentation. The first step is to quantize the colors in an image to several classes. A quantization parameter is involved, which is set to 100 in our experiments. Based on the quantized colors, the second step computes a J value indicating the strength of boundaries and utilizes a region growing method to segment the image based on J values. Users need to specify the number of window sizes used to compute J values. Since aerial images often contain small important objects, we set the number to 4. To alleviate oversegmentation, segments from the second step are merged based on color histograms. We vary the merging threshold from 0.05 to 10.0 to produce a set of segmentations.

The mean-shift approach offers a new tool to solve segmentation. The algorithm iteratively computes mean-shift vectors to map pixels in the joint spatial and color domain to their cluster centers. After convergence, clusters are further merged based on similarity condition. Three parameters are involved—the

spatial bandwidth h_s, the color bandwidth h_r, and the size of the smallest cluster M. Although both the bandwidths can be tuned to reflect the change of analysis scale, we find that for our data set changing M gives more reliable results and shows a smoother tradeoff between precision and recall. We empirically set h_s and h_r to 15 and 10, respectively.

The MSEG algorithm is widely used in the remote sensing community. In MSEG, the increase in heterogeneity when merging a pair of segments is computed as the weighted sum of color and shape heterogeneity measures. We set the weights for color and shape attributes as 0.6 and 0.4, respectively. The merging procedure iteratively merges the pair of segments that results in the least heterogeneity increase until the change in heterogeneity exceeds a set threshold. In our experiments, the threshold varies from 20 to 220.

From a graph point of view, pixels are treated as nodes and their edge weights correspond to the difference in pixel features. In the Felz-Hutt algorithm, a segment corresponds to a connected component in a graph. By defining the differences within a component (*internal difference*) and between two components (*difference between*), the algorithm iteratively merges components whose difference between is smaller than their internal differences. A parameter k in the algorithm represents the observation scale, where a larger value causes the result to have larger segments. We use k values from 100 to 10 000 to obtain segmentations with different scales.

The SRM algorithm utilizes a simple merging procedure coupled with a sorting operation to segment images with great efficiency. Two regions are merged if the mean pixel values of two regions are closer than a merging threshold, which is derived from the perspective of inference problem. Controlling the coarseness of segmentation can be achieved by tuning a parameter Q, which is set within the range of 16–256 in our experiments.

The gPb-owt-ucm algorithm achieves the performance closest to human segmentation so far on the BSD. This approach starts from the contour detector [4], [17] that combines brightness, color, and texture and computes an oriented signal based on spectral graph methods. An oriented watershed transform is applied to the output of contour detector—a probabilistic boundary map to form initial regions, and an ultrametric contour map is constructed to produce a hierarchical segmentation. Segmentations can be produced with different scale values. We use the values in the range of 0–2.

The factorization-based segmentation (FSEG) algorithm first computes the local spectral histogram [18] at each pixel location, which is a concatenated histogram of different filter responses within a local window. Based on the view that each feature can be approximated through linear combination of several representative features and combination weights indicate the region ownership of the corresponding pixel, a feature matrix $\mathbf{Y}$ with columns representing all feature vectors can be expressed as a product of two matrices, $\mathbf{Y} = \mathbf{Z}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$. Here, each column of $\mathbf{Z}$ is a representative feature for a homogeneous region, each column of $\boldsymbol{\beta}$ is the combination weights at each pixel location, and $\boldsymbol{\varepsilon}$ is the noise. A pixel belongs to the region corresponding to the largest weight. The FSEG algorithm utilizes singular value decomposition and nonnegative matrix factorization, which efficiently estimates the factored matrices that immediately give segment labels.

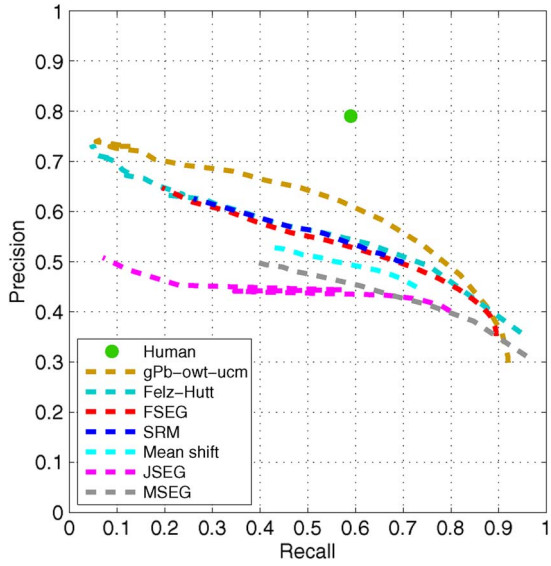

Fig. 3. Precision-recall curves of segmentation algorithms on the aerial image data set.

TABLE I
BOUNDARY BENCHMARKS ON THE AERIAL IMAGE DATASET

	ODS	OIS
Human	0.68	0.69
gPb-owt-ucm	0.62	0.65
Felz-Hutt	0.59	0.62
FSEG	0.58	0.61
SRM	0.58	0.60
Mean shift	0.56	0.58
MSEG	0.53	0.57
JSEG	0.54	0.56

TABLE II
REGION BENCHMARKS ON THE AERIAL IMAGE DATASET

	PRI		VOI	
	ODS	OIS	ODS	OIS
Human	0.83	0.84	1.05	1.00
gPb-owt-ucm	0.62	0.69	1.55	1.52
FSEG	0.62	0.66	1.82	1.81
Felz-Hutt	0.61	0.69	1.47	1.46
JSEG	0.59	0.66	1.65	1.64
SRM	0.58	0.60	2.19	2.18
MSEG	0.49	0.50	2.79	2.79
Mean shift	0.48	0.48	3.71	3.71

For this data set, the algorithm is applied with three filters (the intensity filter and two Laplacian of Gaussian filters) and a small window size (9×9), which produces an over-segmented result. Then, we use a merging procedure to produce hierarchical segmentations. For all connected segments, we iteratively merge the adjacent pair with the weakest common boundary. At a pixel location (x, y), we compute the feature difference between pixel locations at a distance of h along multiple orientations. h is the side length of local window. χ^2-statistics is employed to measure the distance. The boundary strength at (x, y) takes the maximum over orientations. Here, four equally spaced orientations are used. In order to take into account boundaries at multiple scales, we take the sum of boundary strength computed with different window sizes (9×9, 13×13, and 21×21 are used in our experiments). The merging stops when the smallest boundary strength exceeds a threshold, which represents a scale parameter. We change the scale parameter to from 0.1 to 4.5 to produce different segmentations.

Fig. 4. Segmentation results of images in Fig. 1 from seven algorithms. Each row shows the segmentations produced by one algorithm.

IV. Results

Each algorithm is applied to our data set with a set of scale values. By computing the average precision and recall at different scales, a precision-recall curve of each algorithm can be plotted, shown in Fig. 3. A summary score of precision and recall can be given by the F-measure. Table I reports the overall F-measure of each algorithm under two different scale settings. One is the score under optimal data set scale (ODS), where the average F-measure of 80 images at each scale is calculated and the best measure across scales is reported. The other is the score under optimal image scale (OIS), which uses the best F-measure across scales for each image and the average measure over images is reported. These two quantities are used in [12]. The region-based measures, PRI and VOI, are summarized in Table II using the same quantities.

First of all, it can be seen from all the measures that there is a clear gap between the performances of human and algorithms. In Fig. 3, we can see that gPb-owt-ucm outperforms the other algorithms by a noticeable margin, which indicates the better boundary quality. The curves of Felz-Hutt, FSEG, and SRM are largely overlapped with each other. As shown in Table I, the F-measures of the three algorithms are also very close. It should be noted that the segmentations produced by the three algorithms are different in nature, which will be illustrated with examples. For the region-based measures, the ranking of the algorithms is generally similar with that for boundary-based measures, with some exceptions. For example, JSEG is more favored by the region-based measures. The reason is that at large scale levels JSEG often produces several small segments and a large background segment, which result in a low boundary recall but are not well detected by region-based measures.

Fig. 4 displays segmentation results for the six images in Fig. 1, where each segment is randomly colored. For each algorithm, the result corresponding to the best F-measure is presented. Some characteristics of the algorithms can be observed. The algorithms of gPb-owt-ucm and FSEG are able to extract the large objects with complex patterns thanks to the effective use of texture information. One example is that in the fifth image (from left to right) both algorithms produce meaningful segments for the parking lot. A close comparison reveals that the gPb-owt-ucm algorithm does better than FSEG on eliminating noisy regions and retaining main structures. However, they tend to over-smooth the boundaries of small objects, like the airplanes in the fourth image. Although Felz-Hutt and SRM over-segment heavily textured regions, they can preserve the details on boundaries, which result in better segmentation

qualities for small objects, and at the same time handling noisy regions. Between the two algorithms, Felz-Hutt appears to be more sensitive to fine details, while SRM is more effective to capture larger structures. In the mean shift results, the boundaries are well localized, but the oversegmentation problem often occurs due to the simple merging step. Both JSEG and MSEG can produce reasonable results, but the segmentation quality is inferior to others as shown by the examples here.

Computation time is a key concern in many applications. Based on our experiments on an Intel 2.1 GHz machine, Felz-Hutt, SRM, and FSEG are the most efficient algorithms, which take less than 5 s to segment an image. Mean shift, JSEG, and MSEG generally run in 30 s to 1 min for an image. The running time of gPb-owt-ucm is between 7 and 10 min per image. The high computational cost can negatively affect its application to large volume data.

V. Concluding Remarks

We have presented a new aerial image data set with human generated segmentations, which can be used to evaluate segmentation performance on aerial images. Based on this data set, we have conducted a benchmarking study of seven segmentation algorithms. The benchmarking results show that gPb-owt-ucm, Felz-Hutt and FSEG give better performance than the others based on quantitative measures and visual inspection. Despite relatively low region-based measures, the performance of SRM is very competitive, especially considering the efficiency.

A major challenge faced by all the algorithms is their inability to maintain segmentation quality across different scales. As shown in Section IV, the algorithms that perform well for large complex regions tend to smooth out small objects, while the algorithms capable of detecting fine details often over-segment large-scale objects. How to incorporate automatic scale selection to improve segmentation needs to be addressed in further work.

Acknowledgment

This letter has been authored by employees of UT-Battelle, LLC, under contract DE-AC05-00OR22725 with the U.S. Department of Energy. Accordingly, the U.S. Government retains and the publisher, by accepting the article for publication, acknowledges that the U.S. Government retains a non-exclusive, paid-up, irrevocable, world-wide license to publish or reproduce the published form of this manuscript, or allow others to do so, for U.S. Government purposes.

References

[1] 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. ICCV*, 2001, pp. 416–423.

[2] G. Meinel and M. Neubert, "A comparison of segmentation programs for high resolution remote sensing data," *Int. Arch. Photogramm. Remote Sens.*, vol. 35, pp. 1097–1105, 2004.

[3] N. Clinton, A. Holt, J. Scarborough, L. Yan, and P. Gong, "Accuracy assessment measures for object-based image segmentation goodness," *Photogramm. Eng. Remote Sens.*, vol. 76, no. 3, pp. 289–299, Mar. 2010.

[4] D. Martin, C. Fowlkes, and J. Malik, "Learning to detect natural image boundaries using local brightness, color and texture cues," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 5, pp. 530–549, May 2004.

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

[6] M. Meila, "Comparing clusterings: An axiomatic view," in *Proc. ICML*, 2005, pp. 577–584.

[7] 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.

[8] D. Comaniciu, "Mean shift: A robust approach toward feature space analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 5, pp. 603–619, May 2002.

[9] U. C. Benz, P. Hofmann, G. Willhauck, I. Lingenfelder, and M. Heynen, "Multi-resolution, object-oriented fuzzy analysis of remote sensing data for GIS-ready information," *ISPRS J. Photogramm. Remote Sens.*, vol. 58, no. 3/4, pp. 239–258, Jan. 2004.

[10] R. Nock and F. Nielsen, "Statistical region merging," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 11, pp. 1452–1458, Nov. 2004.

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

[12] P. Arbelaez, 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.

[13] J. Yuan and D. L. Wang, "Factorization-based texture segmentation," Dept. Comput. Sci. Eng.,The Ohio State Univ., Columbus, OH, USA, Tech. Rep.OSU-CISRC-1/13-TR01, 2013.

[14] MeanShiftSrc. [Online]. Available: http://coewww.rutgers.edu/riul/research/code/EDISON/

[15] MSEGSrc. [Online]. Available: http://www.berkenviro.com/berkeleyimgseg/

[16] SRMSrc. [Online]. Available: http://www.mathworks.com/matlabcentral/fileexchange/25619-image-segmentation-using-statistical-region-merging

[17] M. Maire, P. Arbelaez, and J. Malik, "Using contours to detect and localize junctions in natural images," in *Proc. CVPR*, 2008, pp. 1–8.

[18] X. Liu and D. L. Wang, "A spectral histogram model for texton modeling and texture discrimination," *Vis. Res.*, vol. 42, no. 23, pp. 2617–2634, Oct. 2002.