International Symposium on Signal Processing and its Applications (ISSPA), Kuala Lumpur, Malaysia, 13 – 16 August, 2001. Organized by the Dept. of Microelectronics and Computer Engineering, UTM, Malaysia and Signal Processing Research Centre, QUT, Australia

A REVIEW OF RECENT EVALUATION METHODS FOR IMAGE SEGMENTATION

Yu Jin ZHANG
Department of Electronic Engineering, Tsinghua University, Beijing 100084, CHINA
zhangyj@ee.tsinghua.edu.cn

ABSTRACT

An up-to-date review of recent progress in the evaluation of image segmentation, since the first survey on this subject published 5 years ago [1], is presented. The analysis and comparison of these evaluation methods are performed according to the classification and assessment criteria for methods and performance metrics proposed in that survey. The results reveal the advantages and limitation of these new methods, and provide additional understanding about the evaluation procedure. This review presents also some novel procedures for image generation under different conditions. Compared the results to that survey, it seems that though more attentions have been attracted and more results have been obtained in these years, new efforts and ideas for the methodology and practical implementation of evaluation are still needed.

1 INTRODUCTION

Objective and quantitative segmentation evaluation plays an important role in image segmentation [2]. The task can be divided into two classes. One is segmentation comparison that is an inter-technique process for ranking the performance of different techniques in segmenting the same type of images. Another is segmentation characterization that is an intra-technique process for recognizing the behavior of the considered technique in segmenting various kinds of image [3].

In 1996, the first extensive survey on the evaluation of image segmentation was published [1]. Since then, works are still going on to propose new evaluation strategies, methods and results. This paper attempts to provide an up-to-date review of recent progress on them.

According to the classification scheme for evaluation methods and criteria proposed in [1], a number of recent evaluation works [4-17] are investigated. Among these works, some of them proposed interesting criteria for performance assessment, some of them suggested useful procedures for generating test images.

In the next section, some important results in [1] are briefly reported. Section 3 gives a short description for each new method and the proposed performance measure. These methods and measures are classified and compared in section 4. Section 5 presents the procedures for simulating various phenomena to generate test images. Special methods are discussed in section 6. Finally, section 7 provides some concluding remarks.

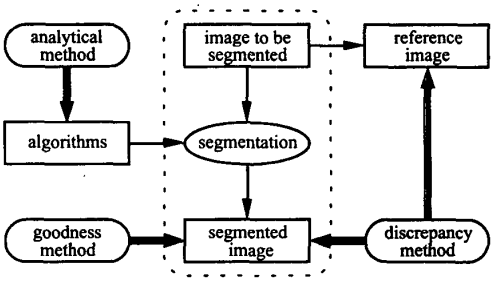

Figure 1. Segmentation and evaluation scheme.

2 RETROSPECTION

We start the study by making a short retrospection on the main results presented in [1], in particular, the general classification of evaluation methods and the list of performance assessing measures (criteria).

A simplified version of the general scheme for segmentation and its evaluation [1] is presented in Figure.1. In Figure 1 the part enclosed by the rounded square with dot line corresponds to the image segmentation procedure which can be considered as a black box of processing with the image to be segmented as the input and the segmented image as the output. Three groups of evaluation methods are distinct: analytical method, (empirical) goodness method and (empirical) discrepancy method. The access points for these three groups of methods are depicted in Figure 1 with thick arrows.

The analytical methods treat the algorithms for segmentation directly by considering the principles, requirements, utilities, complexity, *etc.*, of algorithms. Though it looks simple, the analytical study can not obtain all properties of segmentation algorithms. On the other side, both analysis and practice show that the analytical methods themselves can only provide some additional information to that of other methods do and thus are seldom to be used in isolation.

The goodness methods evaluate the performance of algorithms indirectly by judging the segmented images with certain quality measures established according to human intuition. Various goodness measures have been proposed to cover different aspects of an "ideal" or a "good" segmentation. They have been classified into three groups as shown in Table 1. Note that some of them have also been used in designing segmentation algorithms.

0-7803-6703-0/01/$10.00©2001 IEEE

Table 1. Measures used in empirical evaluation.

#	Method group	Measure group
G-1	Goodness	Intra-region uniformity
G-2	Goodness	Inter-region contrast
G-3	Goodness	Region shape
D-1	Discrepancy	Number of mis-segmented pixels
D-2	Discrepancy	Position of mis-segmented pixels
D-3	Discrepancy	Number of objects in the image
D-4	Discrepancy	Feature values of segmented objects
D-5	Discrepancy	Miscellaneous quantities

The discrepancy methods count the difference between an actually segmented image and a correctly or ideally segmented image (reference image, also called gold standard or ground truth) to assess the performance of segmentation algorithms. In other words, these methods try to determine how far the actually segmented image is from the reference image. Comparative experiments have shown that the discrepancy methods are more effective than the goodness methods [1]. Five measure groups for discrepancy are also listed in Table 1.

3 METHOD AND MEASURE DESCRIPTION

In the following, short descriptions for reviewed methods and their proposed measures are provided.

Borsotti [4] argued that a parameter-free measure would be very useful in evaluation, and designed two enhanced versions of an existing goodness measure (see [3]). This existing measure has incorporated (multiply) two goodness criteria (intra-region uniformity and inter-region contrast) as well as a penalization factor inversely proportional to the number of regions in segmented images (related to D3). The first modified version enhances the contribution of small region for the penalization factor. The second modified version enhances the contribution of the number of small regions and the regions having a large segmentation error.

Rosenberger [5] proposed a genetic technique enable to combine the results of several segmentations. In this technique, adapted to the texture segmentation case, the fusion is based on an evaluation criterion that takes into the consideration of both intra-region homogeneity and inter-region distinction.

Betanzos [6] defined an accuracy measure for the case of segmenting images with multi-types of object. The two main considerations in defining the accuracy measure are (1) workable in cases where not all types of objects are present in each image; (2) able to count the correct and false results separately for each type of object.

Suppose the image contains N types of object, then the accuracy measure is computed by:

$$Accuracy = \sum_{i=1}^{N} \frac{\text{correct segmented pixels in } i\text{th object}}{\text{total number of pixels in } i\text{th object}} \quad (1)$$

A complementary measure for error is also defined.

Chang [7] used two criteria in evaluation, one is a pixel-based measure – mis-classified pixel [1], and another is a region-based measure that is adapted from [8]. In the latter, the degree of overlap between ground-truth and segmented image is considered. Five categories of regions are distinct: (1) correct segmented; (2) over-segmented; (3) under-segmented; (4) missed and (5) noise. The performance measures for the first 4 types of regions are:

$$R_i = \frac{\text{number of regions in } i\text{th category}}{\text{total number of regions in image}} \qquad i = 1,2,3,4 \quad (2)$$

Yang [9] carried out the quantitative assessment of segmentation algorithms for 3-D MRI by calculating a cross-correlation matrix in which each entry stands for the number of voxels identified as object type P by the ith method but identified as object type Q by the jth method. This can be considered as an extension of the confusion matrix method (see [1]). Based on this matrix, the mis-classified percentage of voxels is computed for evaluation.

Zhang [10] presents an objective and quantitative study of different segmentation algorithms using ultimate measurement accuracy (UMA) [18]. This study is distinguished from many other studies by considering (1) both segmentation characterization and comparison; (2) all four types of segmentation algorithms [3]; (3) the use of syntactic images generated in two steps [19].

Mattana [11] made an evaluation of segmentation on the context of check recognition. The evaluation is based on assessing the performance of the segmentation algorithms according to the final character recognition rate. The principle is similar to that of UMA [18]. Though only worked on thresholding techniques by the author, it can be easily extended to evaluate other segmentation algorithms.

Huo [12] incorporated a proposed segmentation algorithm first into an automatic classification scheme, which is composed of three modules: (1) segmentation, (2) feature extraction and (3) classification. Then, the effect of the segmentation algorithms on the performance of the entire scheme is evaluated. Two steps are carried out in the evaluation: (1) computer the area of overlap between segmented regions with expert-identified region; (2) substitute the segmented boundary by expert-identified one, and compare the classification results. The area of overlap for regions A and B is computed as:

$$\text{area of ovelap} = \frac{A \cap B}{A \cup B} \quad (3)$$

It is worth note that by using this goal orientated technique, they got the conclusion with experiments that though the automatically segmented results are worse than the expert-delineated results, it is sufficient for the subsequent feature-extraction and classification tasks to accurately characterize objects.

Xu [13] developed a segmentation algorithm by dividing an image into a number of regions with two traditional conditions in mind, that is, the sum of gray-level variations of these regions is minimized and the average gray-levels between adjacent region is maximized. To reduce the computation time, a tree representation is used. Then, the noise influence on the tree representation and tree partitioning is evaluated using two special measures for performance. First, according to the authors, their algorithm should be effective if a homogeneous region could be represented as one subtree. However, tested with Gaussian noise and transmission noises, they

find "both types of noise have very little effect" on the above property. Second, the influence of noise would affect the result of tree partition by creating a new subtree that has a significantly different average gray level. Similar test results have also been obtained by using the above two models of noise [13].

4 METHOD AND MEASURE COMPARISON

The classification of the above reviewed methods according to the categories of method and measure is given in Table 2.

Table 2. Method and measure groups.

Method #	Source	Group
M-1	Borsotti [4]	G-1, G-2, D-3 like
M-2	Rosenberger [5]	G-1, G-2
M-3	Betanzos [6]	D-1
M-4	Chang [7]	D-1 (modified)
M-5	Hoover [8]	D-1 (modified)
M-6	Yang [9]	D-1
M-7	Zhang [10]	D-4
M-8	Mattana [11]	D-4
M-9	Huo [12]	D-1, D-4
M-10	Xu [13]	D-5

To evaluate different segmentation algorithms, the following four factors are considered for the methods and measures used [1,3]: (1) generality for evaluation; (2) subjective versus objective; (3) complexity for evaluation and (4) evaluation requirements for reference images. A comparison of the methods mentioned in section 3 with these four factors is summarized in Table 3.

Table 3. Comparison of reviewed methods.

Method #	Generality	Sub./Obj.	Complexity	Reference
M-1	Yes[1]	Sub.	Med./High	No
M-2	Yes	Sub.	Med./High	No
M-3	Yes	Obj.	Medium	Yes
M-4	Yes[2]	Obj.	Medium	Yes
M-5	Yes	Obj.	Medium	Yes
M-6	Yes	Obj.	Medium	Yes
M-7	Yes	Obj.	Medium	Yes
M-8	Yes	Obj.	Low/Med.	Yes
M-9	Yes	Obj.	Medium	Yes
M-10	No[3]	Obj.	High	Yes

1: More suitable for being used with test images composed of numerous regions.
2: In conditions the 5 categories of regions can be suitably determined
3: Only usable for algorithms with tree structure

5 IMAGE GENERATION

For applying empirical discrepancy methods, reference images are needed. In recognizing the drawbacks, such as subjectivity and intra- and/or inter-observer variability, in evaluation by comparing automatically segmented results with manually delineated results, synthetic images are adopted in a number of evaluation studies. Synthetic images have the advantages that they can be easily manipulated and reproduced. In addition, ground truth images for synthetic images can be precisely known, which is necessary for a quantitative evaluation [2].

One essential requirement for synthetic images is that the common characteristics of real applications in mind should be reflected [2]. In various domains, quite different images are produced with distinctive procedures.

Xu [13] has generated both additive Gaussian noise and transmission noise for aerial images. To generate transmission noise, a probability P is pre-determined, each pixel in an image has this probability to keep its original gray level during transmission and has the probability $1\text{-}P$ to randomly change to an arbitrary gray level in [0, 255].

Yang [9] has developed a procedure for generating synthetic MRI by using a two-step procedure corresponding to [19]. In the first step, a real MRI is acquired and manually segmented by expert (for incorporating the characteristics of the real MRI) to obtain the basic image. In the second step, noise simulation is made by adding up additive Gaussian noise, while simulation of various degree and different directional in-homogeneity are made by using a sinusoidal based function and the function value is multiplied to the basic image to produce the required test images.

Wagenknecht [14] has also designed a software phantom for evaluation and validation of MRI segmentation algorithms in two steps.

First, different object variations are simulated: (1) to simulate the grayscale distribution, a Gaussian distribution of gray level is used; (2) to simulate the anatomical variability, a Gaussian distribution in space is added at a user-pointed position; (3) to correlate the gray level of neighboring voxel, a Gaussian low-pass filtering is conducted

Second, different artifacts are also simulated: (1) the partial volume effect (caused by the passing of object border inside a voxel) is simulated by using Gaussian low-pass filter; (2) the bias field (to smooth the additional local gradient caused by inhomogeneous sensitivity at the border of different objects) is simulated by multiplying voxel value with a scaling factor that depends on the orientation and dimension of a given voxel neighborhood; (3) the spoiling artifact (can produce band structure across image and change signal intensity and contrast in object space) is generated in form of a triangular profile along trans-axial slice.

6 SPECIAL EVALUATION METHODS

There are also few special evaluation methods proposed in these years.

Dong [15] compared two segmentation algorithms for SAR images with Gaussian distribution and Gamma distribution. Similar to the work of MacAuley (see [1]), only visual comparison is executed for a qualitative evaluation. Visual comparison is also performed in [9].

Gill [16] proposed a semi-automatic segmentation

algorithm. It consists of selecting an inner position in the contour of object for the initialization, then taking a deform model that can rapidly inflate to approximate the contour, and finally using image based forces to better localize the contour. The location of the expert selected initial point can affect the correctness of detected contour, when selected differently the segmented contour would be different. The variability of contour with respect to the location of initial point is an index of algorithm performance and is evaluated by computing the mean and standard deviation.

His colleague Mao [17] further evaluated this semi-automatic segmentation algorithm by using contour probability distribution. The manually outlined contour is used as "truth" contour, its gravity is selected and then used to find the segmented contour by an active contour technique. If the interior is marked as "1" and the exterior is marked as "0", adding all thus obtained binary images can form a gray-scale image, which provides a visual representation showing the variability and accuracy of the segmentation algorithm (this is called intra-object test). The accuracy is computed by the position difference between the mean contour and specific contour (contour under consideration); the variability is determined by the standard deviation.

The above works [16-17] characterized the influence of seed selection on boundary determination, and the principle used for judgement is quite similar to the consensus based approach proposed by Bryant (see [1]).

7 CONCLUDING REMARKS

A number of recent evaluation methods for image segmentation are reviewed. From the analysis and comparison, the following points can be noted:

(1) Image segmentation is the entry to image analysis [3], so the result of image analysis is heavily depends on the quality of segmentation. From the other side, the quality of segmentation should be evaluated in the context of an application and in consideration of the ultimate goal of segmentation [18]. From Table 2, it is clear that in recent years more methods are shifted in this direction.

(2) Though different aspects of segmentation algorithms should be evaluated by using different performance measures, only few methods have used several criteria, especially for discrepancy ones.

(3) Using manually delineated object boundary has the advantages than simple geometric approximation for producing more realistic test images. To make these images representative, the image manually treated should be carefully selected.

(4) Medical image analysis is an important application of general image analysis techniques, it does not only attract many attentions from segmentation development, and it also attracts many attentions from evaluation side. The reason can be multifold, such as the complexity of image contents, the higher precision requirement [20].

(5) Evaluation is not only used in assessing the performance of segmentation algorithms, but now also used to combine the results of several segmentations [5] and to direct the selection of appropriate segmentation algorithms [21].

In conclusion, though more attentions have been attracted, more efforts have been put on and more results have been obtained in the last five years, new ideas and procedures for the methodology and practical implementation of evaluation are still required.

References

[1] Y.J. Zhang. "A survey on evaluation methods for image segmentation." *PR*, 29(8): 1335-1346, 1996.

[2] Y.J. Zhang et al "Objective and quantitative segmentation evaluation and comparison." *Signal Processing*, 39(3): 43-54, 1994.

[3] Y.J. Zhang. *Image Segmentation*, Science Publisher, 2001

[4] M. Borsotti, et al. "Quantitative evaluation of color image segmentation results." *PRL*, 19(8): 741-747, 1998

[5] C. Rosenberger, et al. "Genetic fusion: application to multi-components image segmentation." *Proc. ICASSP*, 4: 2223-2226, 2000

[6] A.A. Betanzos, et al. "Analysis and evaluation of hard and fuzzy clustering segmentation techniques in burned patient images." *IVC*, 18(13): 1045-1054, 2000

[7] K.I. Chang, et al. "Evaluation of texture segmentation algorithms." *Proc. CVPR*, 1: 294-299, 1999

[8] A. Hoover, et al. "An experimental comparison of range image segmentation algorithms." *IEEE PAMI*-18(7): 673-689, 1996

[9] J. Yang, et al. "Method for evaluation of different MRI segmentation approaches." *IEEE NS*-46(6): 2259-2265, 1999

[10] Y.J. Zhang. "Evaluation and Comparison of Different Segmentation Algorithms." *PRL*, 18(10): 963-974, 1997.

[11] M.F. Mattana, et al. "Evaluation by recognition of thresholding-based segmentation techniques on Brazilian bankchecks." *SPIE*, 3572: 344-348, 1999

[12] Z. M. Huo, et al. "Evaluation of an automated segmentation method based on performances of an automated classification method." *SPIE*, 3981: 16-21, 2000

[13] Y. Xu, et al. "A segmentation algorithm for noisy images: design and evaluation." *PRL*, 19(13): 1213-1224, 1998

[14] G. Wagenknecht, et al. "Simulation of 3D MRI brain images for quantitative evaluation of image segmentation algorithms." *SPIE*, 3979: 1074-1085, 2000

[15] Y.H. Dong, et al. "Evaluation of radar image segmentation by Markov random field model with Gaussian distribution and Gamma distribution". *Proc. IGARSS*, 3: 6-10, 1998.

[16] J.D. Gill, et al. "Development and evaluation of a semi-automatic 3D segmentation technique of the carotid arteries from 3D ultrasound images." *SPIE*, 3661: 214-221, 1999

[17] F. Mao, et al. "Technique for evaluation of semi-automatic segmentation methods." *SPIE*, 3661: 1027-1036, 1999

[18] Y.J. Zhang, et al. "Segmentation evaluation using ultimate measurement accuracy." *SPIE 1657*: 449-460, 1992.

[19] Y.J. Zhang, et al. "On the design of test images for segmentation evaluation." *Proc. EUSIPCO-92*, 1: 551-554, 1992.

[20] D.R. Haynor. "Performance evaluation of image processing algorithms in medicine: a clinical perspective." *SPIE*, 3979: 18, 2000

[21] Y.J. Zhang, et al. "Optimal selection of segmentation algorithms based on performance evaluation", *Optical Engineering*, 39(6): 1450-1456, 2000