A Subjective Method for Image Segmentation Evaluation

Qi Wang and Zengfu Wang*

Dept. of Automation, University of Science and Technology of China
crabwq@mail.ustc.edu.cn, zfwang@ustc.edu.cn

Abstract. Image segmentation is an important processing step in many image understanding algorithms and practical vision systems. Various image segmentation algorithms have been proposed and most of them claim their superiority over others. But in fact, no general acceptance has been gained of the goodness of these algorithms. In this paper, we present a subjective method to assess the quality of image segmentation algorithms. Our method involves the collection of a set of images belonging to different categories, optimizing the input parameters for each algorithm, conducting visual evaluation experiments and analyzing the final results. We outline the framework through an evaluation of four state-of-the-art image segmentation algorithms—mean-shift segmentation, JSEG, efficient graph based segmentation and statistical region merging, and give a detailed comparison of their different aspects.

Keywords: Image segmentation, subjective evaluation.

1 Introduction

Image segmentation is an important processing step in many image, video and computer vision applications. Extensive research has been done in creating many different approaches and algorithms for image segmentation [1-10]. However, no single segmentation technique is universally useful for all applications and different techniques are not equally suited for a particular task. Hence there needs a way of comparing them so that the better ones can be selected. To properly position the state of the art of image segmentation algorithms, many efforts have been spent on the development of performance evaluation methods.

Typically, researchers show their segmentation results on a few images and point out why their results look better than others. In fact, we never know from such studies if their results are good or typical examples, whether they are for a particular image or set of images, or more generally, for a whole class of images. Other evaluation methods include analytical and empirical goodness methods [11]. For analytical methods [12, 13], performance is judged not on the output of the segmentation method but on

* Corresponding author: Zengfu Wang. Address: Dept. of Automation, University of Science and Technology of China, Hefei, Anhui 230027, P.R.China. This work was supported by the National Natural Science Foundation of China (No.60875026).

H. Zha, R.-i. Taniguchi, and S. Maybank (Eds.): ACCV 2009, Part III, LNCS 5996, pp. 53–64, 2010.

the basis of their properties, principles, complexity, requirements and so forth, without reference to a concrete implementation of the algorithm or test data. But until now, this kind of methods may only be useful for simple algorithms or straightforward segmentation problems. The difficulty is the lack of general theory for image segmentation [14]. As for empirical goodness methods, some goodness metrics such as uniformity within regions [15], contrast between regions [16] and shape of segmented regions [17] are calculated to measure the quality of an algorithm. The great disadvantage is that the goodness metrics are at best heuristics, and may exhibit strong bias towards a particular algorithm [18]. To address these problems, it has been widely agreed that a benchmark, which includes a large set of test images and some objective performance measures, is necessary for image segmentation evaluation. Several important works [19-23] emerged and among these, one widely influential prior work is Berkeley benchmark presented by Martin et al. [19]. Unfortunately, both Martin's and other researchers' work suffer from a series of shortcomings, which are discussed in [20].

This paper presents a segmentation evaluation method that was motivated by the following two proposals.

(1) The first one is that an evaluation method should produce results that correlate with the perceived quality of segmentation images. This was noted by Cinque et al. [24]:"Although it would be nice to have a quantitative evaluation of performance given by an analytical expression, or more visually by means of a table or graph, we must remember that the final evaluator is man and that his subjective criteria depend on his practical requirements." Though those methods mentioned above can be very useful in some applications, their results do not necessarily coincide with the human perception of the goodness of segmentation.

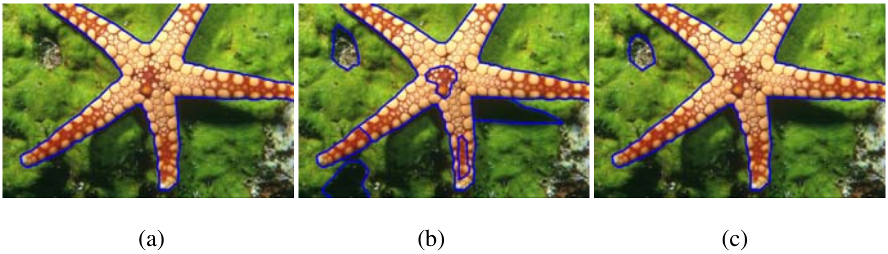

(a) (b) (c)

Fig. 1. Illustration of segmentation comparison where the blue boundaries in the images indicate the segmentation results. The left image (a) is the ground-truth segmentation. The middle (b) and right (c) are respectively the results produced by two segmentation algorithms.

(2) The second one is that existing benchmark based evaluation methods, usually objective methods, cannot properly reflect the goodness of different segmentation algorithms, so human subjects are needed to directly evaluate the output of segmentation algorithms. Generally, these methods define different functions, which measure the discrepancy between an algorithm's results and the ground-truth segmentations, to produce a quantitative value as a representation of the algorithm's quality. But actually, the human labeled ground-truth segmentations are another kind of expression of the pictures' semantic meaning and it is not convincing to measure it by a quantitative

value, especially when there are several ground-truth segmentations for one picture just as Martin's benchmark. Indeed, it is still a known difficult task to exactly quantify the semantic meaning. "In the absence of explicit semantics, the only alternative is to turn to human subjects, who will introduce implicit semantics through their understanding of the images [23]." Take the segmentations in Fig. 1 as an example. Suppose the left image (a) is a ground-truth segmentation, the middle (b) and right (c) are two segmentation results by different algorithms. If we evaluate the algorithms through hit rate by comparing the segmented boundaries with the ground-truth ones, the two algorithms will be considered as the same in performance. Nevertheless, most of us will think the algorithm producing the right segmentation result (c) is better than the one producing the middle one (b).

The approach taken to evaluate segmentation algorithms in this work is to measure their performance by human subjects, to use real images of different types in the evaluation and to select the parameters for each algorithm in a meaningful way that is not biased towards any algorithm. Aspects that distinguish our work with the previous are the following:

(1) Firstly, we test segmentation algorithms on images of different types and analyze their performance separately. This is often overlooked by other evaluation methods, which usually draw a thorough conclusion on a bunch of mixed test images. As a matter of fact, it can be a distinguished property that different algorithms may perform differently on each categorized images.

(2) Secondly, our selected input parameters of each algorithm for the final evaluation process are more reasonable. Most exiting evaluation methods merely gave a mathematic metric without considering the parameter selection problem or challenged this crucial step with ambiguity. In this work, we use a coarse-to-fine method to select 10 "best" parameter sets for each algorithm from a large parameter space. The final evaluation is made on the basis of each algorithm's 10 parameter sets. Therefore, our conclusion is more robust and can reflect the algorithms' real performance.

(3) In our analysis of the experimental data, we use the statistical technique and psychological model Intraclass Correlation Coefficient. This makes our experimental conclusion more reasonable and acceptable.

The remainder of this paper is organized as follows. In Section 2, we introduce the images used in our experiments and four algorithms to be evaluated. In Section 3, we describe our parameter selection procedure for each algorithm and in Section 4, experiments are conducted to assess the performance of the four segmentation algorithms. Finally, discussion and conclusion are made in Section 5.

2 Images and Segmentation Algorithms

Any scheme for evaluating segmentation algorithms must choose a test-bed of images with which to work. In this paper we employ the publicly available Berkeley image segmentation database [25], to which existing evaluation method frequently refer. 50 natural images of different types are carefully selected from the database. They are categorized as textured and nontextured, each of which compose half of the dataset. To ensure wide variety, we intentionally collect images with various contents, such as human, animal, vehicle, building, landscape, etc. All images are colored RGB format of 481×321 or 321×481 in size.

We also select four segmentation algorithms in our evaluation, which are mean-shift segmentation (MS) [4], JSEG [1], efficient graph-based method (EGB) [3] and statistical region merging (SRM) [2], based on the following three considerations.

(1) They well represent different categories of image segmentation methods.

(2) All of them are relatively new methods and published in well-known publications.

(3) The implementations of these methods are publicly available.

3 Parameter Selection

Selecting the input parameters of each algorithm is a critical step in performance evaluation because the resulting quality varies greatly with the choice of parameters. Most existing evaluation methods treat this complex problem with ambiguity or do not mention it at all. In this paper, we select parameters in a prudent manner. For each algorithm, the plausible meaningful range of each parameter is determined by consulting the original paper and doing a preliminary experiment, through which we can get a general idea of the parameters' effects on the algorithm's results. We try our best to make sure each parameter of a specific algorithm samples the entire reason-able parameter space, with no bias toward any parameter or algorithm. After this initial parameter selection, we then choose ten parameter settings for each algorithm through five persons' evaluation. Our final results are based on the ten parameter settings of each algorithm. This is called the final parameter selection.

3.1 Initial Parameter Selection

According to the principles mentioned above, we choose the initial combinations of parameter settings for each algorithm as follows.

(1) Mean-shift segmentation (MS). The mean-shift based segmentation technique is one of many techniques under the heading of "feature space analysis." There are three parameters for the user to specify. The first parameter h_s, and second parameter h_r, are respectively the radius of the spatial dimensions and color dimensions for gradient estimation. The third one, M (minimum region), controls the number of regions in the segmented image. Our preliminary experiment on dozens of images tells us that the reasonable maximum of the three parameters are respectively about 49, 30.5 and 7000. Therefore, we give 7×7×7 combinations of mean-shift parameters, where $h_s \in \{7, 14, 21, 28, 35, 42, 49\}$, $h_r \in \{6.5, 10.5, 14.5, 18.5, 22.5, 26.5, 30.5\}$ and $M \in \{50, 200, 700, 1000, 3000, 5000, 7000\}$.

(2) JSEG segmentation (JSEG). JSEG is a much more different method based on region growing using multiscale "J-images." The algorithm has three parameters that need to be determined by the user. The first one is a threshold q for the quantization process. The second one is the region merging threshold m and the last one l is the number of scales desired for the image. The ranges of the three parameters are bounded by the author in the implementation as $q \in [0, 600]$, $m \in [0, 1]$ and $l \in \{1, 2, 3\}$. Consequently, the initial 7×7×3 JSEG parameter settings are combinations of

the just referred three, where $q \in \{85, 170, 255, 340, 425, 510, 595\}$, $m \in \{0.15, 0.30, 0.45, 0.60, 0.75, 0.90, 1.00\}$ and $l \in \{1, 2, 3\}$.

(3) Efficient graph-based segmentation (EGB). This is typically a graph-based segmentation method by comparing and merging pairwise regions. The algorithm required three parameters to be set: σ , k and Min . σ is used to smooth the input image before segmenting it. k is the value for the threshold function and Min represents the minimum component size enforced by post-processing. On the basis of a preliminary experiment, the initial $7 \times 7 \times 7$ parameter settings are determined as $\sigma \in \{0.15, 0.30, 0.45, 0.60, 0.75, 0.90, 1.00\}$, $k \in \{200, 500, 800, 1000, 2000, 3000, 4000\}$ and $Min \in \{50, 200, 700, 1000, 3000, 5000, 7000\}$.

(4) Statistical region merging based segmentation (SRM). The key idea of this method is to formulate image segmentation as an inference problem and then process it with region merging and statistical means. There is only one parameter Q , which control the coarseness of the segmentation, to be decided by the user. Q is an integer number confined in the range of [1, 256] according to the original paper but our preliminary experiment shows a shrunken range of [1, 80] is more appropriate.

We can easily find that all the first three algorithms (MS, JSEG and EGB) have three parameters and we ensure each parameter equally samples the reasonable parameter space. Some of the parameters have the same meaning (e.g. M and Min). Thus they are given the same numerical value. While this method does not guarantee that the optimal input parameter set is identified — indeed there is currently no accepted method that will guarantee finding the optimal input parameters without ground truth — it does avoid biasing the results toward any of the algorithm. Unfortunately, the fourth SMR algorithm only depends on one parameter. This makes it harder to compare it with the other three. Adding two more parameters by modifying the algorithm is a way of addressing this problem [22, 26]. However, this is not an easy task and furthermore, modifying the algorithm may divert it greatly from the original one. Thereby we handle this demanding problem by shrinking the initial parameter settings to less than one third of the first three parameters' choices.

3.2 Final Parameter Selection

In this stage, the number of every algorithm's initial parameter combinations is reduced to 10. The methodology employed here is by the subjective evaluation of 5 participants, major in computer vision, on 20 images. Half of the images are textured and the other half are nontextured.

In the first place, each algorithm produces results on the 20 training images with all its initial parameter combinations. But the forms of segmentation results differ greatly with each other as showed in Fig. 2. MS produces results with two kinds of forms, white-black boundary map and region map with mean color within the region. JSEG gives its segmentation results in the form of boundary map superimposed on the original map, while EGB with region map filled with random color and SRM with boundary map superimposed on the mean color region map. In order to exempt the influence of different segmentation representations on participants' ratings, we program to transform the three kinds of results of MS, EGB and SRM into one uniform representation, boundary map superimposed on the original map, which is the same as the results of JSEG. Participants are then asked to rate the segmentation results from a scale of one to seven. There is no time limit but the participants are asked to make their standard of "goodness" consistent in the whole procedure and for one image, the scores of four algorithm must be rated at one time. The rating score indicates the easiness of identifying the perceptually different objects from the segmentation results. The higher the score is, the easier the different objects can be identified. After doing this, the results of each algorithm with a rating no less than five are collected in together. We then use a voting process to decide the most representative parameter settings for each algorithm. Parameters with the highest ten voting scores are selected as the algorithm's final parameter settings. The results of final parameter selection are listed in Table 1.

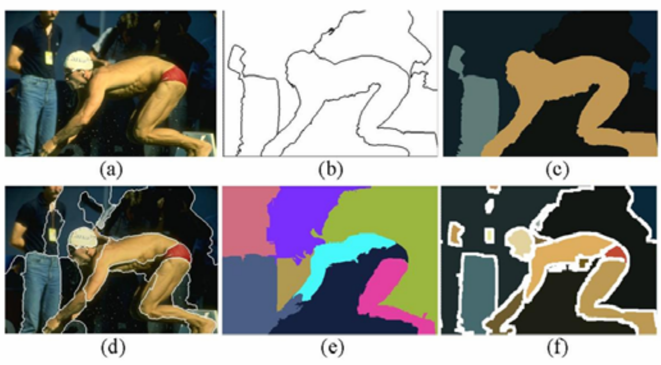


Fig. 2. The different representations of the four algorithms' segmentation results: (a) an original image; (b, c) the segmentation results of MS; (d) the segmentation result of JSEG; (e) the segmentation result of EGB; (f) the segmentation result of SRM

4 Algorithm Evaluation

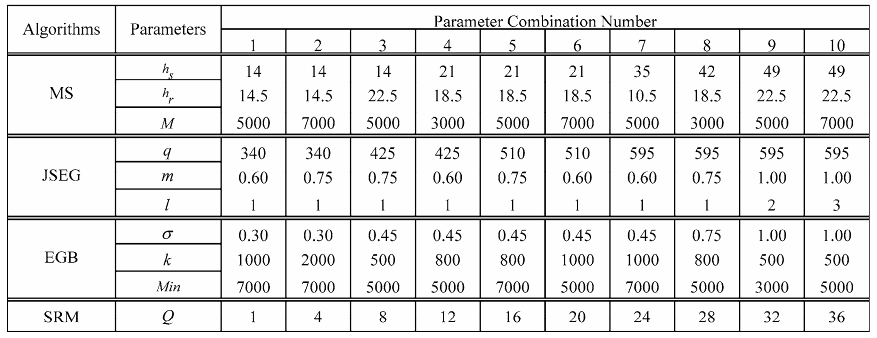
Getting the ten parameter settings for each algorithm, we then use a subjective way to evaluate the four segmentation algorithms. In this experiment, 30 images are used. Each image is processed by each algorithm with all its 10 parameter combinations. The total 1200 segmentation results are then evaluated by 20 persons major in computer vision in a similar way as described in the final parameter selection step.

4.1 Consistency between Participants' Ratings

Before we start our evaluation, it is important to known whether the ratings are consistent across the participants. This is estimated using one form of the Intraclass Correlation Coefficient psychological model [27, 28]. The ICC (3, k) form is appropriate for the task because it measures the expected consistency of the k participants' mean ratings. The ICC (3, k) model is defined as:

$$\mathrm{ICC}(3, \mathrm{k}) = \frac{\mathrm{bms - ems}}{\mathrm{ems}}, \tag{1}$$

Table 1. The final parameter selection results. There are 10 parameter combinations for each algorithm.



Algorithms	Parameters	Parameter Combination Number									
		1	2	3	4	5	6	7	8	9	10
MS	h_s	14	14	14	21	21	21	35	42	49	49
	h_r	14.5	14.5	22.5	18.5	18.5	18.5	10.5	18.5	22.5	22.5
	M	5000	7000	5000	3000	5000	7000	5000	3000	5000	7000
JSEG	q	340	340	425	425	510	510	595	595	595	595
	m	0.60	0.75	0.75	0.60	0.75	0.60	0.60	0.75	1.00	1.00
	l	1	1	1	1	1	1	1	1	2	3
EGB	σ	0.30	0.30	0.45	0.45	0.45	0.45	0.45	0.75	1.00	1.00
	k	1000	2000	500	800	800	1000	1000	800	500	500
	Min	7000	7000	5000	5000	7000	5000	7000	5000	3000	5000
SRM	Q	1	4	8	12	16	20	24	28	32	36

where bms represents mean square of the ratings between targets, ems means total error mean square, and k is the number of participants. The values of ICC can range from zero to one, where zero means no consistency and one means complete consistency. This ICC model is used for every segmentation algorithm to examine the correlation of participants' ratings. Since 20 participants are involved in the procedure, the ICC (3, 20) for each algorithm is 0.9258(MS), 0.9702(JSEG), 0.9692(EGB) and 0.9617 (SRM). These figures give an indication that a consensus about ratings exists. This is a critical result since it establishes the validity of comparing the ratings in our experiments.

4.2 Performance and Parameters

In this part, we examine the algorithms' performance under different parameters. In the first place, the rating scores are used to determine the two parameter settings using two different criteria. The best single overall parameter setting, termed the fixed parameters, is identified by averaging the ratings across the participants, averaging these results across images, and finding the parameter set with the highest average. The best parameter setting for each individual image, termed the adapted parameters, is also found. This is done by averaging the ratings across participants and identifying the parameters that have the highest average rating for each image.

The fixed parameters for each algorithm are: MS (35, 10.5, 5000), JSEG (595, 0.60, 1), EGB (1.00, 500, 3000) and SRM (32). The adapted parameters for each image are show in Table 2. Since all the images are from Berkeley database, their names are labeled with numbers as they were.

We calculate their mean ratings under the two parameter settings by averaging the ratings across images and participants and then compare their relative performance. For fixed parameters, their mean ratings are 4.73(MS), 4.76(JSEG), 4.37(EGB) and 4.13(SRM). For adapted ones, they are 5.12(MS), 5.01(JSEG), 4.73(EGB) and 4.43(SRM).We can see clearly that the performance of the two algorithms — MS and JSEG — are better than that of EGB and SRM in both fixed and adapted parameter settings, while for MS and JSEG, or EGB and SRM, the difference in them is trivial. We can also find that for every algorithm, the adapted parameters outperform the

Table 2. The adapted parameters for each image. Image names are labeled with the numbers in the Berkley database.

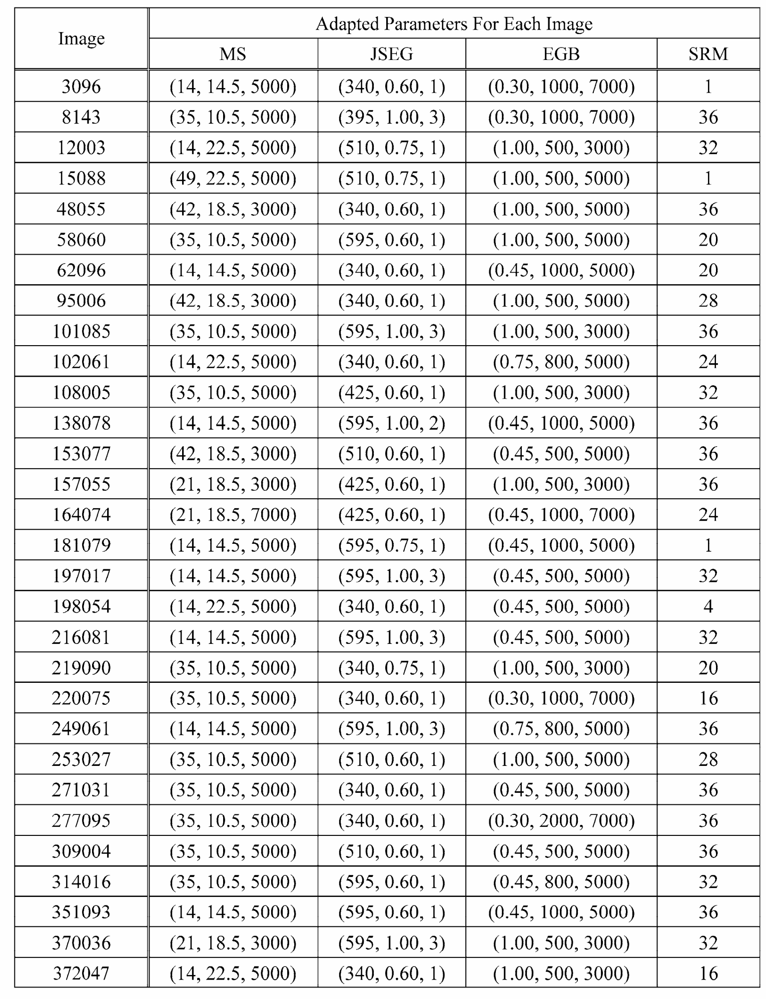


Image	Adapted Parameters For Each Image			
	MS	JSEG	EGB	SRM
3096	(14, 14.5, 5000)	(340, 0.60, 1)	(0.30, 1000, 7000)	1
8143	(35, 10.5, 5000)	(395, 1.00, 3)	(0.30, 1000, 7000)	36
12003	(14, 22.5, 5000)	(510, 0.75, 1)	(1.00, 500, 3000)	32
15088	(49, 22.5, 5000)	(510, 0.75, 1)	(1.00, 500, 5000)	1
48055	(42, 18.5, 3000)	(340, 0.60, 1)	(1.00, 500, 5000)	36
58060	(35, 10.5, 5000)	(595, 0.60, 1)	(1.00, 500, 5000)	20
62096	(14, 14.5, 5000)	(340, 0.60, 1)	(0.45, 1000, 5000)	20
95006	(42, 18.5, 3000)	(340, 0.60, 1)	(1.00, 500, 5000)	28
101085	(35, 10.5, 5000)	(595, 1.00, 3)	(1.00, 500, 3000)	36
102061	(14, 22.5, 5000)	(340, 0.60, 1)	(0.75, 800, 5000)	24
108005	(35, 10.5, 5000)	(425, 0.60, 1)	(1.00, 500, 3000)	32
138078	(14, 14.5, 5000)	(595, 1.00, 2)	(0.45, 1000, 5000)	36
153077	(42, 18.5, 3000)	(510, 0.60, 1)	(0.45, 500, 5000)	36
157055	(21, 18.5, 3000)	(425, 0.60, 1)	(1.00, 500, 3000)	36
164074	(21, 18.5, 7000)	(425, 0.60, 1)	(0.45, 1000, 7000)	24
181079	(14, 14.5, 5000)	(595, 0.75, 1)	(0.45, 1000, 5000)	1
197017	(14, 14.5, 5000)	(595, 1.00, 3)	(0.45, 500, 5000)	32
198054	(14, 22.5, 5000)	(340, 0.60, 1)	(0.45, 500, 5000)	4
216081	(14, 14.5, 5000)	(595, 1.00, 3)	(0.45, 500, 5000)	32
219090	(35, 10.5, 5000)	(340, 0.75, 1)	(1.00, 500, 3000)	20
220075	(35, 10.5, 5000)	(340, 0.60, 1)	(0.30, 1000, 7000)	16
249061	(14, 14.5, 5000)	(595, 1.00, 3)	(0.75, 800, 5000)	36
253027	(35, 10.5, 5000)	(510, 0.60, 1)	(1.00, 500, 5000)	28
271031	(35, 10.5, 5000)	(340, 0.60, 1)	(0.45, 500, 5000)	36
277095	(35, 10.5, 5000)	(340, 0.60, 1)	(0.30, 2000, 7000)	36
309004	(35, 10.5, 5000)	(510, 0.60, 1)	(0.45, 500, 5000)	36
314016	(35, 10.5, 5000)	(595, 0.60, 1)	(0.45, 800, 5000)	32
351093	(14, 14.5, 5000)	(595, 0.60, 1)	(0.45, 1000, 5000)	36
370036	(21, 18.5, 3000)	(595, 1.00, 3)	(1.00, 500, 3000)	32
372047	(14, 22.5, 5000)	(340, 0.60, 1)	(1.00, 500, 3000)	16

fixed parameters. This is a significant result because it implies that the amount of effort expected in parameter optimization can influence the measured performance of the algorithm. Therefore, equal effort must be put in optimizing the parameters in a real application.

4.3 Performance and Image Category

In this experiment, we examine the interaction between the algorithms' performance and the image categories. The relative performance of the algorithms is calculated separately by averaging the ratings across images of a specific category and participants. For textured images, their mean ratings are 4.28(MS), 3.78(JSEG), 3.66(EGB) and 3.03(SRM). For nontextured ones, they are 4.56(MS), 4.76(JSEG), 4.21(EGB) and 4.32(SRM). Generally, the four algorithms perform better on nontextured images than on textured images. This suggests that each algorithm leaves something to be improved when confronted with textured images. How to deal with texture is still a problem required to be noticed while developing segmentation algorithms. As for their relative performance, we can find that the MS algorithm performs significantly better than the other three for textured images, while SRM is the poorest and, JSEG and EGB are nearly of the same level. For nontextured images, JSEG and MS produce better results than SRM and EGB. The difference between MS and JSEG is marginal. The same is true for EGB and SRM.

4.4 Stability with Respect to Different Images

Performance variation with respect to different images under one particular parameter combination is an important property. Here we first average ratings of every image under every parameter setting across participants and then calculate the variance of 30 images' ratings under every parameter combination. At last the 10 variances of the algorithm are averaged as a representation of the algorithm's stability under different images. The variances for the four algorithms are respectively 2.85(MS), 2.55(JSEG), 1.38(EGB) and 1.79(SRM). From these figures, we can see that the stability of the four algorithm is EGB>SRM>JSEG>MS. Though MS and JSEG produce better results than EGB and SRM, their stability with respect to images is not as good as EGB and SRM.

4.5 Stability with Respect to Different Parameter Settings

An algorithm's performance may vary greatly under different parameter settings. In this experiment, we look at an algorithm's stability across the 10 best parameter combinations. We average the ratings of a particular image across participants and then compute the variance of an image's 10 ratings. After that, we average the variance results across 30 images. Experimental results are 0.91(MS), 1.45(JSEG), 0.46(EGB) and 1.16(SRM), which show that the relative stability are EGB>MS>SRM>JSEG. When employing a sensitive algorithm such as JSEG, we should pay more attention to the parameter selection because it may affect the results greatly.

4.6 Processing Time Comparison

Processing speed is a critical consideration in many applications. Sometimes it is much more important than other properties discussed above. However, except for MS algorithm, none of the other three algorithms give a running time registration in their original implementation programs. So we make a little change in the original programs to make them capable of registering the processing time while segmenting an image. After that, we calculate the mean processing time of an algorithm by averaging time across all images and parameter settings. All the programs are run on a computer with Pentium 4 CPU 2.93GHz and 1G memory.

The processing time (in seconds) are 44.34(MS), 9.88(JSEG), 0.61(EGB) and 0.39(SRM). Obviously, we can see that MS is the most time-consuming algorithm, so it is not appropriate for real time applications. JSEG runs more quickly than MS, but it still can not satisfy the need of real time situations. Fortunately, EGB and SRM are both quick enough in a real time system and SRM stands out with the quickest speed.

5 Discussion and Conclusion

In this paper, we have presented a subjective method for comparing the quality of image segmentation algorithms. To demonstrate the utility of our proposed method, we performed a detailed comparison between four algorithms: mean-shift segmentation (MS), JSEG, efficient graph-based method (EGB) and statistical region merging (SRM). The algorithms were compared with respect to different parameter settings, image categories and processing time. Also, two kinds of stability were considered: stability with respect to parameters for a given image and stability with respect to different images for a given parameter combination. Our experimental results show that no single algorithm can outperform others in all aspects mentioned above. For example, MS and JSEG perform better than EGB and SRM in terms of parameter settings and different image categories, while their stability and processing time are not as good as the other two properties. Therefore, there should be a trade-off between these characteristics in the selection of a real application.

We can also find that, from the perspective of recognizing the different objects in images, even the state-of-the-art segmentation algorithms are far from perfect. This can be demonstrated from the mean scores in Section 4.2 and Section 4.3. We believe that, only after knowing how to solve this object recognition segmentation, can we make a great progress in image segmentation. Additionally, an effective object recognition segmentation method can facilitate many related applications, such as contend based image retrieval. Our future research involves developing a new segmentation algorithm consistent with human perception and this work is under way.

Our comparison in this experiment is an overall one rarely done in previous evaluation papers. We can get a complete understanding of the algorithms after this evaluation. This is informative when confronting with a problem of segmentation method selection in real applications.

However, this evaluation method has its shortcomings. First, subjective evaluation is a tedious and time-consuming work. In these experiments, the entire 50 images require thousands of ratings for every participant. This severely limits the number of images used in the evaluation, which brings out the second shortcoming that the ability to generalize our experimental results may be limited. In this work, 20 images were used in the parameter selection process and another 30 images are used in the algorithm evaluation process. This is not a large number compared with those objective evaluation methods. In spite of this, we argue that our evaluation conclusion is meaningful and useful. For one reason, the rating scores of different participants are consistent with each other as the psychological model ICC (3, k) demonstrates. For another, though it is not a large number of images, they have diverse image characteristics, and it is larger enough than those which claim their superiority over others on just several images. Besides, some of the properties compared in our experiments vary greatly with different algorithms and we believe it can reflect the actual quality of the algorithm.

References

1. Deng, Y., Manjunath, J.B.S.: Unsupervised Segmentation of Color-texture Regions in Images and Video. IEEE Transactions on Pattern Analysis and Machine Intelligence 23(8), 800–810 (2001)
2. Nock, R., Nielsen, F.: Statistical Region Merging. IEEE Transactions on Pattern Analysis and Machine Intelligence 26(11), 1452–1458 (2004)
3. Felzenszwalb, P.F., Huttenlocher, D.P.: Efficient Graph-based Image Segmentation. International Journal of Computer Vision 59(2), 167–181 (2004)
4. Comaniciu, D., Meer, P.: Mean shift: a Robust Approach toward Feature Space Analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence 24(5), 603–619 (2002)
5. Shi, J., Malik, J.: Normalized Cuts and Image Segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence 22(8), 888–905 (2000)
6. Cheng, H.D., Jiang, X.H., Wang, J.: Color Image Segmentation Based on Homogram Thresholding and Region Merging. Pattern Recognition 35, 373–393 (2002)
7. Crevier, D.: Image Segmentation Algorithm Development Using Ground Truth Image Data Sets. Computer Vision and Image Understanding 112(2), 143–159 (2008)
8. Benlamri, R., Al-Marzooqi, Y.: Free-form Object Segmentation and Representation from Registered Range and Color Images. Image and Vision Computing 22, 703–717 (2004)
9. Mushrif, M.M., Ray, A.K.: Color Image Segmentation: Rough-set Theoretic Approach. Pattern Recognition Letters 29, 483–493 (2008)
10. Wang, S., Siskind, J.M.: Image Segmentation with Ratio Cut. IEEE Transactions on Pattern Analysis and Machine Intelligence 25(6), 675–690 (2003)
11. Zhang, Y.J.: A Survey on Evaluation Methods for Image Segmentation. Pattern Recognition 29(8), 1335–1346 (1996)
12. Liedtke, C.E., Gahm, T., Kappei, F., Aeikens, B.: Segmentation of Microscopic Cell Scenes. Analytical and Quantitative Cytology and Histology 9(3), 197–211 (1987)
13. Abdou, I.E., Pratt, W.K.: Quantitative Design and Evaluation of Enhancement/thresholding Edge Detectors. Proceedings of the IEEE 67(5), 753–763 (1979)
14. Haralick, R.M., Shapiro, L.G.: Computer and Robot Vision. Addison-Wesley, New York (1992)
15. Huang, Q., Dom, B.: Quantitative Methods of Evaluating Image Segmentation. In: International Conference on Image Processing, vol. 3, pp. 53–56 (1995)
16. Levine, M.D., Nazif, A.M.: Dynamic Measurement of Computer Generated Image Segmentations. IEEE Transactions on Pattern Analysis and Machine Intelligence 7(2), 155–164 (1985)
17. Sahoo, P.K., Soltani, S., Wong, A.K.C., Chen, Y.C.: A Survey of Thresholding Techniques. Computer Vision, Graphics, and Image Processing 41(2), 233–260 (1988)
18. Monteiro, F.C., Campilho, A.C.: Performance Evaluation of Image Segmentation. In: Campilho, A., Kamel, M.S. (eds.) ICIAR 2006. LNCS, vol. 4141, pp. 248–259. Springer, Heidelberg (2006)
19. Martin, D., Fowlkes, C., Tal, D., Malik, J.: A Database of Human Segmented Natural Images and Its Application to Evaluating Segmentation Algorithms and Measuring Ecological Statistics. In: Proc. International Conference on Computer Vision, vol. 2, pp. 416–423 (2001)
20. Ge, F., Wang, S., Liu, T.: Image-segmentation Evaluation from the Perspective of Salient Object Extraction. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 1, pp. 1146–1153 (2006)

21. Zhang, H., Fritts, J.E., Goldman, S.A.: Image Segmentation Evaluation: A Survey of Unsupervised Methods. Computer Vision and Image Understanding 110, 260–280 (2008)
22. Unnikrishnan, R., Pantofaru, C., Hebert, M.: Toward Objective Evaluation of Image Segmentation Algorithms. IEEE Transactions on Pattern Analysis and Machine Intelligence 29(6), 929–944 (2007)
23. Shaffrey, C.W., Jermyn, I.H., Kingsbury, N.G.: Phychovisual Evaluation of Image Segmentation Algorithms. In: Proceedings of Advanced Concepts for Intelligent Vision Systems (2002)
24. Cinque, C., Guerra, C., Levialdi, S.: Reply: On the Paper by R. M. Haralick. CVGIP: Image Understanding 60(2), 250–252 (1994)
25. http://www.eecs.berkeley.edu/Research/Projects/CS/vision/grouping/segbench/
26. Health, M.D., Sarkar, S., Sanocki, T., Bowyer, K.W.: A Robust Visual Method for Assessing the Relative Performance of Edge-detection Algorithms. IEEE Transactions on Pattern Analysis and Machine Intelligence 19(12), 1338–1359 (1997)
27. Shrout, P.E., Fleiss, J.L.: Intraclass Correlation: Uses in Assessing Rater Reliability. Psychology Bulletin 86(2), 420–428 (1979)
28. http://www.nyu.edu/its/statistics/Docs/intracls.html