On the Evaluation of Image Segmentation Algorithms

Brendan McCane
Department of Computer Science
University of Otago
Dunedin, New Zealand.
Phone: +64 3 479 8588. Email: mccane@cs.otago.ac.nz

October 1, 1997

Abstract

In this paper, I will address the problems of segmentation evaluation in the context of what types of evaluation can and should be performed. Also, I propose a graphical method for presenting results of empirical evaluation experiments which utilises an independent ordering of data curves for comparison of segmentation algorithms.

Keywords: feature detection, segmentation evaluation, 2D segmentation

1 Introduction

Image segmentation is a fundamental process in many computer vision applications. In fact, image interpretation applications are just complex image segmentation algorithms! In a more conventional sense, however, image segmentation is the partitioning of an image into its constituent regions, where parts within a region are similar according to some uniformity predicate, and dissimilar between neighbouring regions. Due to its importance, many segmentation algorithms have been proposed, and a number of evaluation criteria have also been proposed. In spite of this, very few comparative results of segmentation algorithms have been performed. Hoover et al [1] state that "important areas of computer vision suffer from a lack of sound experimental work". They go on to show that in the area of range image segmentation, there is almost no objective evaluation of results and add that "the situation is characteristic of essentially all of computer vision". This is a worrying trend in an area that has existed for more than 30 years. Many researchers prefer to rely on qualitative human judgement of results for evaluation. In fact, Pal and Pal [4] go so far as to say that "a human being is the best judge to evaluate the output of any segmentation algorithm". I must respectfully disagree with this statement, which I will return to in Section 2.

Hoover et al [1] present several factors for consideration to alleviate this problem, one of which I will attempt to address in this paper states "the comparative framework (for evaluation) is itself a research issue, and so deserves appropriate conceptual energy in its development". Also, Zhang [7] states that "the performance evaluation of segmentation algorithms is indispensable and thus an important subject in the study of segmentation". Although numerous evaluation criteria have been proposed, no-one (as far as I am aware) has explicitly addressed the problem of presenting the results of comparing algorithms. In this paper, I will address the problems of segmentation evaluation in the context of what types of evaluation should be performed, and suggest that the computer vision community adopt a graphical method for presenting results which orders each curve independently for comparison (see Section 3).

2 Evaluation of Segmentation Algorithms

Zhang [7] partitions the evaluation of image segmentation algorithms into three broad categories: analytical, empirical goodness (EG) and empirical discrepancy (ED). While I do not question the usefulness of the analytical approach for evaluation, in this paper I will concentrate on empirical approaches. In empirical goodness measures, the resultant segmentation is evaluated based on some intuitive measure of goodness such as uniform grey-level within regions, or contrast of grey-level between regions ([2]). Empirical discrepancy methods, however, make use of "ground truth" images to determine if each pixel has been classified correctly or not, and there are several methods for evaluating these misclassified pixels. Zhang goes on to show that ED is significantly better than EG for evaluating segmentation algorithms as it has a larger dynamic range of results. However, he uses only a simple image in which ground truth is trivial. Hoover et al ([1]) also assert that single measures of goodness of segmentations should be avoided in favour of ED techniques. While I agree in principle with these statements, they rarely apply to realistic intensity or colour images. Hoover et al even state that "ground truth based on intensity is considerably more subjective than that based on geometry". I would make a stronger claim and say that ground truth based on intensity or colour, in many cases, simply does not exist. Any segmentation of such images is necessarily subjective. In such situations, subjective evaluation of segmentation using EG methods is necessary for evaluation to occur at all. In a more general sense, though, EG techniques can be useful in helping researchers looking at different problems determine if a given technique may be applicable, whereas ED techniques are restricted to the problem domain they were tested on.

An evaluation method which was not mentioned by either Zhang or Hoover et al, and is related to ED measures, involves evaluation based entirely on the application for which the algorithm was designed. In many applications there is an obvious measure for evaluating the segmentation algorithm - how well does the application perform using a particular segmentation algorithm? For example, in object recognition, it is typical to use a segmentation algorithm (which could be edge or region based) to partition the image into a number of parts which are then used for object recognition. Rather than "a human being is the best judge to evaluate the output of any segmentation algorithm" as claimed by Pal and Pal [4], in this case, the application is the best judge of any segmentation algorithm. If the application performs better using a particular algorithm, then that algorithm is better for that application regardless of what a human expert thinks about the quality of segmentation.

To conclude this section, the evaluation of segmentation algorithms is not a simple process and there is no "one true way". I believe that all algorithms should be evaluated using a variety of forms including analytical measures, empirical discrepancy techniques, empirical goodness techniques, application dependent methods, and subjective human visual perception, if the techniques are possible to apply in a given situation (analytical and empirical discrepancy techniques are not always possible or feasible to use). Using as many techniques as possible (but not indiscriminately), will provide a broader and fairer evaluation and comparison of segmentation algorithms.

3 Presentation of Empirical Evaluation Results

Given that evaluating segmentation algorithms using empirical techniques is indispensable, the problem of presenting such results becomes apparent. Firstly, it is useful to identify the characteristics of a good evaluation experiment. The most important characteristics are:

1. The comparison of the current algorithm against several well known techniques.

2. Evaluation of each technique using many images (on the order of hundreds) from many application domains of varying de-

grees of complexity (with emphasis on the current domain).

3. The use of several evaluation measures for comparison.

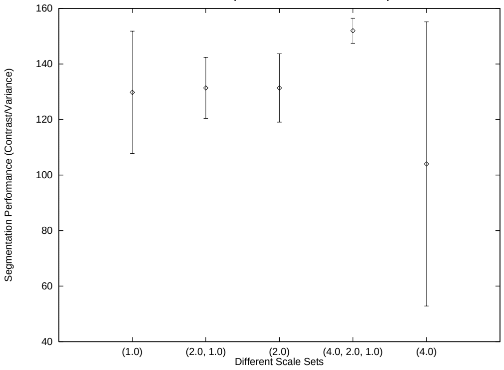
With these points in mind, we have, at the least, several hundred evaluation points (one for each image for each technique for each measure) which need to be presented. Also, given that the test images are of varying degrees of complexity, the variance of these results will be large.

There are a number of well known ways of presenting such data, and some not so well known ways which will be evaluated here. To that end, I will use an evaluation of a segmentation algorithm described in McCane and Caelli ([3]) as a case study. In these papers, a multiscale algorithm is evaluated against its single scale equivalents. The test images used in this evaluation cover a broad cross-section of images including outdoor scenes, indoor scenes, computer generated images, medical images (CT-scans and X-Rays), and one human subject image (Lenna). In total 39 images were used for evaluation and one of the EG measures involved a combined contrast and variance measure which will be used here.

An obvious (and common) method for presentation of results involves using tabular presentation of data. However, this will only work with very small numbers of images. Even with a relatively small number of images (eg 39) tabular presentation becomes impractical. Hence graphical methods are the only sensible techniques for displaying this kind of data.

An obvious graphical method for presenting the data is as an error bar plot. The means and standard deviation of the performance of each algorithm with respect to the EG measure is plotted. However, if numerous images of differing complexity are used for testing, it is likely that the algorithms performance will vary widely depending on the input image, producing large error bars. For example, Figure 1 shows such a plot, using the data presented in McCane and Caelli ([3]). In the figure, the error-bars indicate the standard deviation of the results for each algorithm. As can be seen from the figure, due to the large variances, it is difficult to interpret the results other than to say that, on average, method 5 performs better than the other methods. Of course, this can be a useful result, but it doesn't indicate how well each technique performed on individual images.

Figure 1 (Error-bars:)



Error-bars: some useful information can be obtained from the graph, but this style of presentation for this type of data is not optimal.

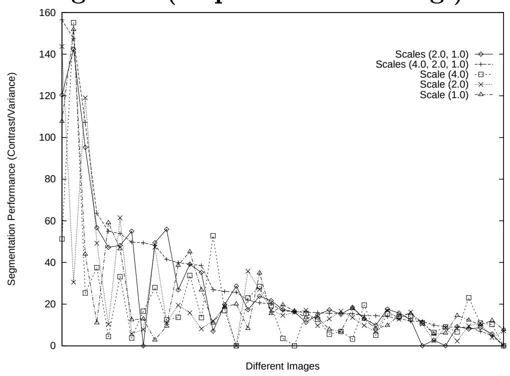
Perhaps a more appropriate method of presenting the data is to have each image represented as a point along the x-axis of a graph, and the EG value is then plotted along the y-axis. The family of points belonging to one algorithm can then be joined to form a curve indicating how that algorithm is performing. Clearly, the order of the images along the x-axis is important since a random ordering will result in a wildly varying curve which will be undecipherable. The question is what ordering should be placed on the x-axis? This same problem has occurred in the area of theorem proving in the automated reasoning community ([6]) and the same graphing techniques can be applied here. There are two obvious options:

1. Order the x-axis such that the algorithm with the best average score has a monotonically increasing or decreasing curve, and plot all results on the same graph.

2. Order each algorithm's curve independently such that all curves are monotonically increasing or decreasing, and again plot all curves on the same graph.

Both methods have their advantages. In the first method, the graph is simpler than the random case, and we can still directly compare the performance of each algorithm on specific images by comparing curves at a specific x-coordinate. The disadvantage here is that all algorithms are ordered according to the best on average, and other curves may be wildly varying in comparison as shown in Figure 2. As can be seen from the figure, the results are difficult to interpret.

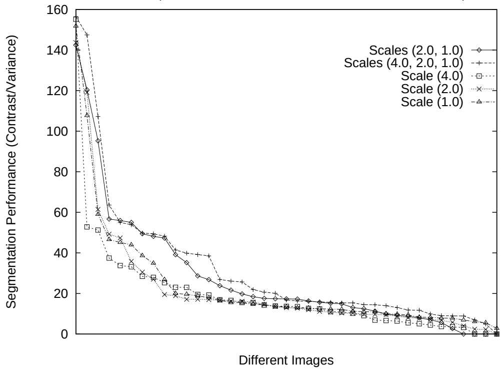
Figure 2 (Dependent ordering:)



Dependent ordering: the graph is confusing due to the effect of the varying curves. The ordered curve is well presented however (Scale set (4.0, 2.0, 1.0))

The alternative method is to order each of the curves independently such that all curves are monotonically increasing or decreasing. This has the advantage of significantly reducing the complexity of the resulting graph and making it easier to interpret. However, it does mean that a direct comparison between particular algorithms cannot be performed at a single x-coordinate in the graph. The resulting graph is not only aesthetically more pleasing, however, but is also intuitively a good comparison since each algorithm is presented according to its own perception of success with respect to the EG measure. So, if one algorithm finds it easy to segment a particular type of image, then it will be compared with the images a competing algorithm segments well. This is analogous to allowing students to choose the questions which they will be examined on (presumably they will choose questions they are good at). Figure 3 shows the results of presenting the same data from Figures 1 and 2 in this manner. The figure clearly shows which algorithm is performing better (Scale set (4.0, 2.0, 1.0)) according to the EG measure and also gives a good indication of how the other algorithms are performing compared to each other which is something the previous two presentation methods fail to do.

Figure 3 (Independent ordering:)



Independent ordering: the graph shows clearly and concisely which algorithms are performing better than the others.

4 The Need for Application-based Results

The data presented in the previous section displays clearly that the multi-scale approach using the scale set $(4.0, 2.0, 1.0)$ outperforms the other scale sets. However, when evaluated according to the application for which a particular algorithm is to be used, the results do not necessarily correlate with the EG measures. For example the recognition results using the different scale sets is presented in Figure 4 ([3]). In this case it makes more sense to present the average performance of the segmentation algorithms rather than the independent ordering method of the previous section, since all images are from the same domain (simple indoor scenes). From this figure it is clear that the optimal scale set is not, in fact, $(4.0, 2.0, 1.0)$ as suggested by the EG measure. But, in conjunction with these results and the results according to the EG measure, the best scale set to choose would probably be $(2.0, 1.0)$ since it has good results using both measures. It is still important to include other measures such as EG

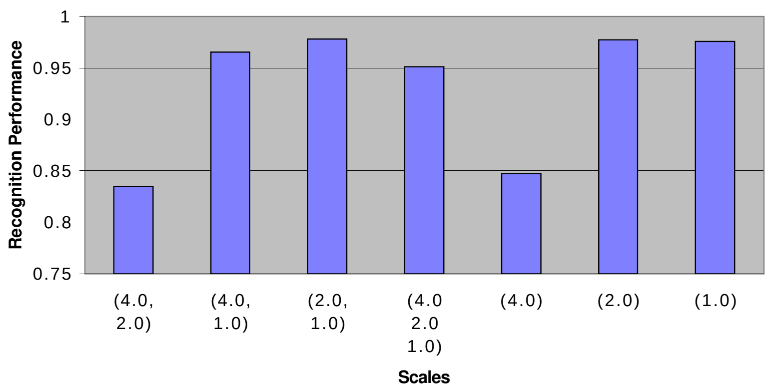


Figure 4: Presentation of recognition results : the graph shows clearly that the scale set $(4.0, 2.0, 1.0)$ is not optimal according to recognition performance.

measures, as application-based evaluation typically only gives results for a specific class of images and does not indicate the general capabilities of the algorithms (which other researchers in different application areas may be interested in).

5 Conclusion

In this paper I have discussed the problems inherent in evaluating segmentation algorithms and have discussed methods for presenting data of empirical evaluation experiments. In particular, whilst I concede that evaluation algorithms based on empirical discrepancy measures is the preferred method, it is seldom possible in the area of segmenting intensity or colour images, since any segmentation of such images is necessarily subjective. Consequently, empirical goodness measures are necessary, but not sufficient, for many evaluation experiments. In addition to EG methods, application-specific measures should also be applied (which do not necessarily fall within the experimental discrepancy domain), to give a balanced evaluation of results. In terms of presenting data from empirical experiments, I have shown that plotting multiple curves based on independent ordering of data on the same graph is the most intuitively and aesthetically pleasing method, although the standard method of plotting means and standard deviations can also be helpful.

Finally, I would like to emphasize the need for the image segmentation community, and the computer vision community as a whole, to develop standards for evaluating and comparing old and new algorithms alike. Towards this end, there needs to be a central repository of algorithm implementations, data, and evaluation measures so that researchers can quickly and effectively compare their novel algorithms with well established algorithms, thus allowing for a rapid maturation of the image segmentation and computer vision field. A repository such as the “Thousands of Problems for Theorem Provers” (TPTP) library [5] would be a good model to follow.

References

[1] Adam Hoover, Gillian Jean-Baptiste, Xiaoyi Jiang, Patrick J. Flynn, Horst Bunke, Dmitry B. Goldgof, Kevin Bowyer, David W. Eggert, Andrew Fitzgibbon, and Robert B. Fisher. An experimental comparison of range image segmentation algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18(7):673, 1996.

[2] Martin D. Levine and Ahmed M. Nazif. Dynamic measurement of computer generated image segmentations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-7:155–164, 1985.

[3] Brendan McCane and Terry Caelli. Multi-scale adaptive segmentation using edge and region based attributes. In *First International Conference on Knowledge-based Intelligent Electronic Systems*, pages 72–80, Adelaide, South Australia, 1997. IEEE.

[4] Nikhil R. Pal and Sankar P. Pal. A review on image segmentation techniques. *Pattern Recognition*, 26(9):1277–1293, 1993.

[5] G. Sutcliffe and C.B. Suttner. The TPTP problem library for automated theorem proving. `http://www.cs.jcu.edu.au/~tptp/`.

[6] G. Sutcliffe and C.B. Suttner. The Results of the CADE-13 ATP System Competition. *Journal of Automated Reasoning*, To appear, 1997.

[7] Y.J. Zhang. A survey on evaluation methods for image segmentation. *Pattern Recognition*, 29(8):1335–1346, 1996.