PARAMETER-ORIENTATED SEGMENTATION ALGORITHM EVALUATION

Hassan Al-Muhairi, Martin Fleury, and Adrian F. Clark

University of Essex, United Kingdom
{hoalmu, fleum, alien}@essex.ac.uk

ABSTRACT

Quantitative testing of segmentation algorithms implies rigorous testing against ground-truth segmentations. Though underreported in the literature, the performance of a segmentation algorithm depends on the choice of input parameters across core, pre- and post-processing stages. The paper highlights the importance of post-processing parameters when the figure of merit is the Berkeley F-measure. It also shows that the search of a parameter space with a genetic algorithm is not only accelerated through the inclusion of a time factor in the cost function but the relative importance of different parameters is highlighted.

Index Terms— Genetic algorithm, Image segmentation, Quantitative testing

1. INTRODUCTION

Quantitative algorithm evaluation is a way to systematically evaluate image segmentation algorithms. In this paper, we take this a stage further by highlighting the role of parameter selection in evaluation. Choice of algorithm parameters can critically determine the final segmentation output for the class of algorithms now considered, those that need to be tested off-line [1]. The paper's contributions are two-fold. Firstly, when hand-segmented images form the ground truth, the importance of parameters that influence the post-processing stage of common segmentation algorithms is established. Secondly, the search for suitable parameters is abbreviated by means of a genetic algorithm in which computation time is factored into the cost function.

Implicit in quantitative testing is that the finite set of test data is representative of variation in the real world and that the number of samples is large enough [2]. At the University of Berkeley, CA, Martin [3] supervised the creation of a database of 12,000 hand-labeled (from a pool of 30 human subjects) segmentations of images taken from the Corel dataset of 1,000 images. The Berkeley database encourages users to download benchmarking code as well as 200 training images and a further 100 test images of size 240×160 pixels. The originators of the database ran a multiple cue segmentation algorithm [3] of their own using this database as a basis for quantitative evaluation. The Berkeley F-measure [4], effectively summarized comparisons with other algorithms, at least in their ability to reproduce human segmentations. This F-measure is the basis of our parameter-wise evaluations and as such is further explained in Section 2.

There are two qualifications when employing quantitative evaluation: the first is that the choice of ground truth can bias the selection of algorithm and its parameters, as further discussed in Section 3; and the second is the time taken for evaluation if an exhaustive search over parameter space is conducted. We have used a genetic algorithm (GA) search module in our evaluation environment (though not for the F-measure experiments) to decrease the processing time for the search as a whole.

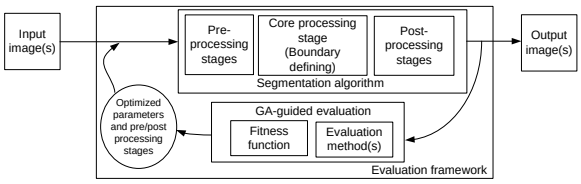

Figure 1. Combined segmentation and GA-guided evaluation framework

The operation of the GA establishes the relevance of parameter settings on the final segmentation quality and its execution time. Furthermore, there is a link between these parameters and the underlying components of the image segmentation algorithm. The process basically is divided into three stages: 1) a pre-processing stage; 2) a core segmentation stage; and 3) a post-processing stage. Though the auxiliary components of the algorithm are well-known, such as smoothing filters, color quantization, and region pruning, they have a significant effect on the segmentation output quality. In fact, it maybe possible to harmonize all segmentation algorithms so that they match this structure. Fig. 1 illustrates an abstract diagram for a future system of combined segmentation and GA evaluation framework.

Section 2 now describes our methodology. Section 3 reports results. Finally, Section 4 concludes the paper.

2. METHODOLOGY

In supervised evaluation of segmentation, ground-truth can be represented by boundary pixels. To compare segmentation algorithm output with ground truth output an error measure is required. Let S be a segmentation result for a single image. Then for segmentations S_1 and S_2 an error measure can be defined [5] as:

$$E(S_1, S_2, p_i) = \frac{|R(S_1, p_i) \backslash R(S_2, P_i)|}{|R(S_1, P_i)|} \tag{1}$$

where $R(S_j, p_i)$ is the region in segmentation j that contains pixel p_i, $\backslash$ denotes the set difference and $|x|$ is the cardinality of set x. The Local Consistency Error (LCE) is one way to summarize the error across all image pixels indexed by i, taking into account the two directions of comparison between segmentations S_1 and S_2:

$$LCE(S_1, S_2) = \frac{1}{n} \sum_i \min(E(S_1, S_2, p_i), E(S_2, S_1, p_i)) \tag{2}$$

A precision-recall data point records the balance between accurate identification of boundary pixels versus the number of boundary pixels detected. Precision is the ratio of true positives

(correctly detected boundary pixels) to false positives, whereas recall is the ratio of true positives detected to the total number of true positives. For a set of precision-recall evaluations of image segmentations, the Berkeley F-measure is a summary statistic that identifies how many misinterpretations (of boundary pixels) will occur to achieve its identification rate. The F-measure is represented as a harmonic mean, weighted by a desired trade-off between precision and recall. Without prior knowledge, this cost α is set as 0.5 in (3):

$$F = PR/(\alpha R + (1-\alpha)P) \tag{3}$$

The maximum value of F represents the optimal achievable value for a given algorithm and a given image type or application. In practice, an exact match between detected boundary pixels and ground-truth pixels is unlikely. One reason for this is the physical variation in the positioning of hand-segmented boundary lines by the human subjects. In fact, the median (not maximum) human F-measure is 0.79 for the 100 images from the Berkeley database. Therefore, a correspondence matching algorithm is required and in this work we adopt the method given in Appendix B of [3].

We tested 'exact' and 'fast' versions of the supplied Berkeley code and found respectively a variation of 3.5–4 hours to 50–60 mins. to test 100 images with just one parameter set on a desktop PC. Specifically, the tests were run on an Intel Core 2 Duo running at a nominal clock speed of 2.66 GHz with 4 GB RAM. A clean installation of Linux Obuntu v. 9.04 had been performed immediately before running the tests. As minimal variation in the resulting F-measure was observed, our results are presented using the fast version.

The GA employed was a real-valued GA inspired by Polheim's GEATbx, a GA toolbox for Matlab. [6] (though the output may be integer-valued parameter settings). It employs extended intermediate recombination [7], whereby an offspring gene g_O is conceived from its two parent genes g_1 and g_2 by

$$g_O = \alpha g_1 + (1-\alpha) g_2 \tag{4}$$

where α is a scaling factor in the range $[-d, 1+d]$. Assuming that the N genes in a chromosome form an N-dimensional hypercube then, when $d = 0$, g_O lies along the straight line between g_1 and g_2; if $d > 0$, extrapolation is permitted, though in this work, $d = 0$. The algorithm allows mutation of variable parameter values as a means of preventing chromosomes becoming too close to each other and remaining in local minima. To avoid the need for a an arbitrary number of generations and the creation of a stopping condition it is possible to refine the results from a fixed number of generations. This was achieved through a non-GA polishing algorithm to what is a non-constrained, non-linear optimization problem. Newtonian methods are unsuitable if the cost function to be optimized is non-differentiable.

Therefore, this work used the Nelder-Mead direct search, 'simplex method' [8]. In the implementation, the final values found by the GA and the cost function form the input to the algorithm. The values form the vertices of the simplex and at each iteration the worst one of these values is replaced. This is achieved by a number of trial evaluations of the cost function at the vertex after the simplex has been reflected, expanded or contracted. For a related combination of GA with hill-climbing algorithm in adaptive image segmentation refer to [9].

3. EXPERIMENTS

3.1. Processing stage parameters, the *F* measure

Three parameters were selected for graph-based segmentation [10] and 18 parameter sets were tested. Fig. 2 is arranged to show the trend that increases the F-measure and precision. of the parameters, `sigma` is the smoothing stage parameter, it is a pre-processing stage. `Threshold`, as the name suggests, is the value for the thresholding stage. `MinimumSize` is minimum size of the regions enforced by the post-processing pruning stage. Three parameters were also selected for the Color Watershed algorithm [11] evaluation, namely: `minProp` — the minimum probability for a gradient value to be relevant to the the watershed; mergeThreshold — values of this parameter close to zero will produce over-segmentation as all gradient values will be detected as edges and values close to one will produce under-segmentations, as few very sharp gradient will be detected as edges; `kernelSize` — the kernel size of the median filter used to de-noise the image (if the value is equal to one then no pre-smoothing will be carried out). The results for 27 Color Watershed parameter sets are plotted in Fig. 3.

There is a similarity between the effect of the Graph-based `MinimumSize` parameter and the Color Watershed `mergeThreshold` parameter. All the parameters can more or less be considered as defining a post-processing pruning stage. The `MinimumSize` parameter depends on the segments' sizes and the other two parameters depend on the similarity of the mean colors of the joined regions. In all cases, increasing the value of these parameters (*i.e.* making the processing stage related to this parameter more 'effective' in the final result) increases the precision values (and decreases the Recall value slightly). As a consequence, the F-measure value is increased. This leads to a nice conclusion about the importance of the post-processing stage that prunes the segmentation results independently of the type of the algorithms used. On the other hand, it is important to note that this outcome reflects the choice of hand-segmented ground truths. As humans usually focus on the principal objects in the middle of images and disregard other part of the image, ground-truth images are usually under-segmented.

3.2. Time factor optimization of GA parameter selection

The mean-shift algorithm [12] makes a convenient example. For these experiments, which involved only the error function, equation (2) and not the F-measure, the population size was set to 20 and the first 20 generations were run. The recombination rate was fixed at 0.6 and the mutation rate at 0.2 (refer to Section 2). Timing appears in the GA cost function as a multiplicative factor applied to the combined evaluation factor. Other ways of including this factor such as by an additive or exponential weighting were found to be ineffective. For example, including timing as an argument of an exponential function forces the GA to optimize heavily to reach a solution with low processing time values.

Fig. 4 shows the application of the GA with and without a time factor in the cost function. The horizontal axis is annotated with the image numbers of 20 images from the Berkeley database. Consider the effect of the time factor on the value of the `radiusS` parameter: when the time factor is present, the value of this parameter is always equal or less than two. While without the presence of the time factor the same parameter value does not have a specific trend, and changes between different images in the test. The best explanation for this is that this parameter does not have a great significance for the quality of the segmentation. However, higher values of this parameter are computationally expensive. There is no similar trend for the `RadiusS` parameter, and the time factor also does have any noticeable effect on the third `ColorDistance` parameter.

By observation, the GA module reaches an acceptable stable solution in much fewer generations when a comparatively large population size is employed. If each evaluation is a lengthy process, then rapid convergence of the GA is anyway desirable. To see the start of the stabilization trend in the parameter search

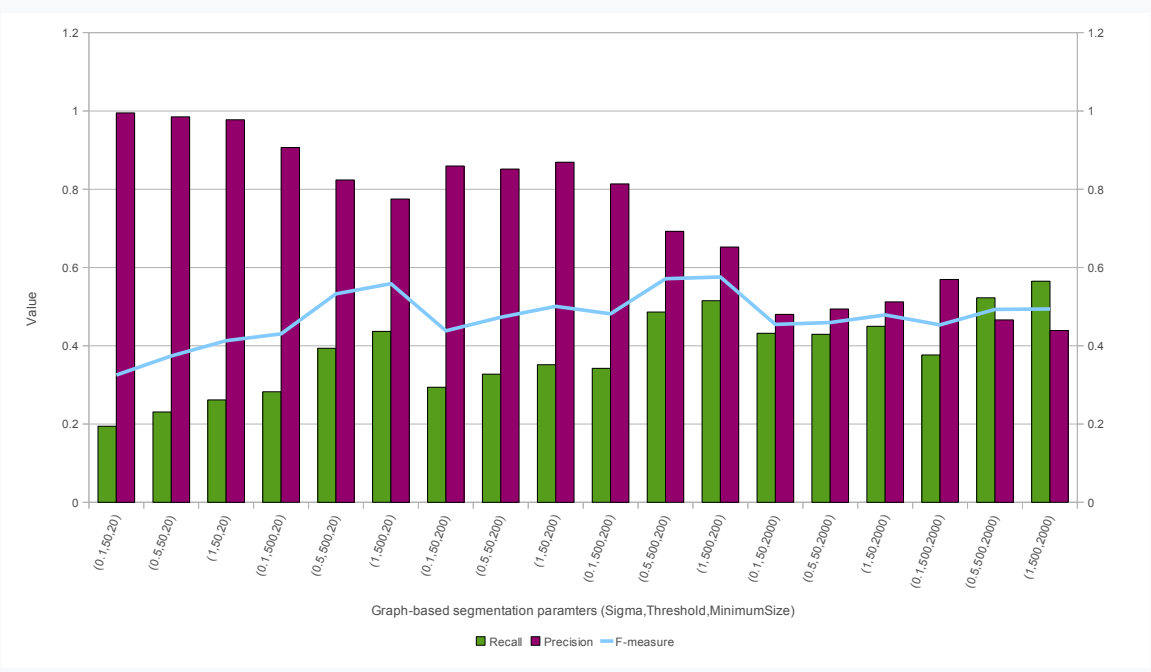

Figure 2. Graph-based algorithm [10]: Maximum F-measure with P and R values for a selection from 100 images and their hand-segmented ground truths from the Berkeley database

with the time factor, Fig. 5 shows the first three generations. The members of the population are plotted across the horizontal axis and the parameter values for a population member can be read off in the vertical direction. At the crossover point between the generations, the fittest parameter set is shown. It is clear from the parameter variation in the second and the third generations that the selection is already stabilizing. For example `RadiusS` tends to stabilize at value 2 and `RadiusR` at value 5. However, it is important to notice that there is still a possibility that this trend can change if a better parameter combination is found in the coming generations. After 20 generations it was found that `RadiusS` was actually lowered to value 1 and `RadiusR`'s value increased to 8, while the `ColourDistance` parameter stabilised at a value of 10 after ten generations.

Therefore, employing a time factor arrives at similar results for the example image but may well increase the convergence speed as the values of less significant parameters are explored less. The evaluation of the 100 images from the Berkeley database without a time factor took 995 mins. = 16.58 hrs, whereas the addition of the time factor reduced the computation by almost a half, i.e. to 523 mins. = 8.72 hrs.

4. CONCLUSIONS

The Berkeley F-measure is perhaps the most appropriate metric for quantitative evaluation of image segmentation algorithms. Yet, this paper has shown that if hand-segmented test images form the ground truth, then evaluation is strongly affected, in two common algorithms, by parameter selection in the post-processing stage. Future work should confirm this finding by harmonizing the processing stages between different segmentation algorithms, so that pre- and post-processing stages are the same and the core segmentation stages are varied.

A GA was deployed to optimize image segmentation parameter selection during the evaluation process. It was found that including a time factor into the GA cost function has three potential advantages: 1) It can increase the convergence speed because less important parameters are not explored in such detail; 2) It arrives at computationally efficient parameter sets; and 3) If GA parameter settings with and without the time factor are examined, insignificant parameters can be isolated.

5. REFERENCES

[1] H. Zhang and J. E. Fritts and S. A. Goldman, "A co-evaluation framework for improving segmentation evaluation," in *SPIE Conf.*, vol. 5809, 2005, pp. 420–430.

[2] I. Guyon and J. Markhoul and R. Schwartz and V. Vapnik, "What size test set gives good error rate estimates?," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 1, pp. 52–64, 1998.

[3] D. R. Martin and C. Fowles 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, 2004.

[4] C. Van Rijsbergen, *Information Retrieval*, Dept. of Computer Science, Univ. of Glasgow, second edition, 1979.

[5] D. Martin and C. Fowles and D. Tal and J. Malik, "A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics," in *Int'l Conf. on Computer Vision*, 2001, pp. 416–423.

[6] H. Polheim, *GEATbxsurvey: Evolutionary algorithm toolbox for MATLAB, version 3.7*, 2005, Online at `http://www.geatbx.com/docu/index.html`.

[7] H. Mühlenbein and D. Schlierkamp-Voosen, "Predictive models for the breeder genetic algorithm: I. continuous parameter optimization," *Evolutionary Computation*, vol. 1, pp. 25–49, 1993.

[8] J. A. Nelder and R. Mead, "A simplex method for function minimization," *Computer Journal*, vol. 7, pp. 308–313, 1965.

[9] B. Bahnu and S. Lee and S. Das, "Adaptive image segmentation using multi-objective evaluation and hybrid search methods,", in *AAAI Fall Symposium*, 1993, pp.30–34.

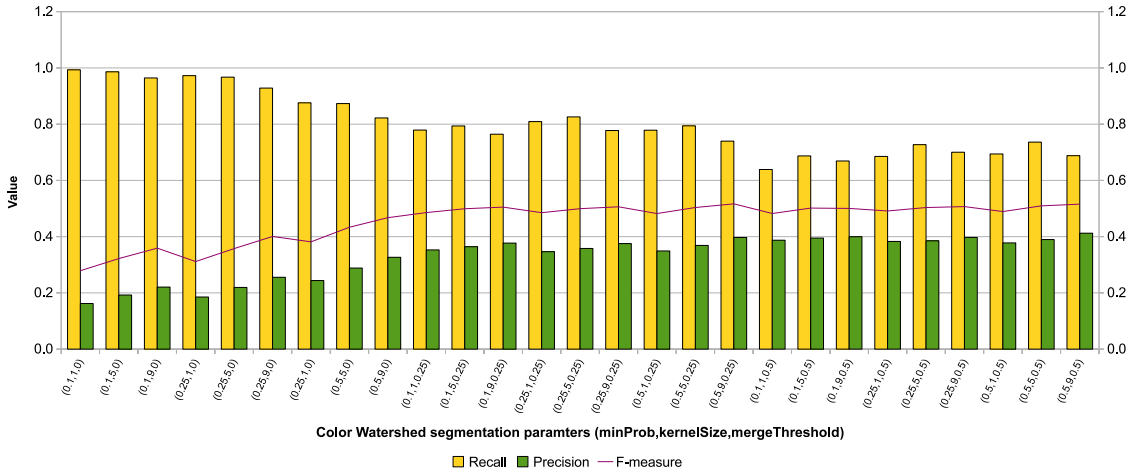

Figure 3. Color watershed algorithm [11] : Maximum F-measure with P and R values for a selection from 100 images and their hand-segmented ground truths from the Berkeley database

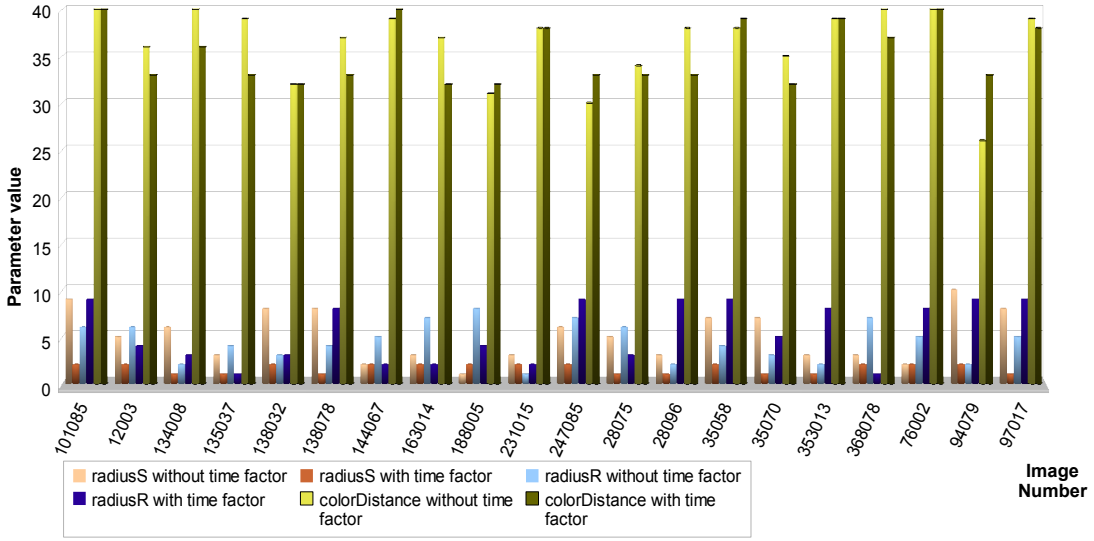

Figure 4. Meanshift segmentation [12] evaluation for 20 images with and without a time factor

[10] P. F. Felzenszwalb and D. Huttenlocher, "Efficient graph-based image segmentation," *Int'l J. of Computer Vision*, vol. 29, no. 2, pp. 167–181, 2004.

[11] L. Vincent and P. Soille, "Watersheds in digital spaces: An efficient algorithm based in immersion simulations," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 13, no. 6, pp. 583–598, 1991.

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

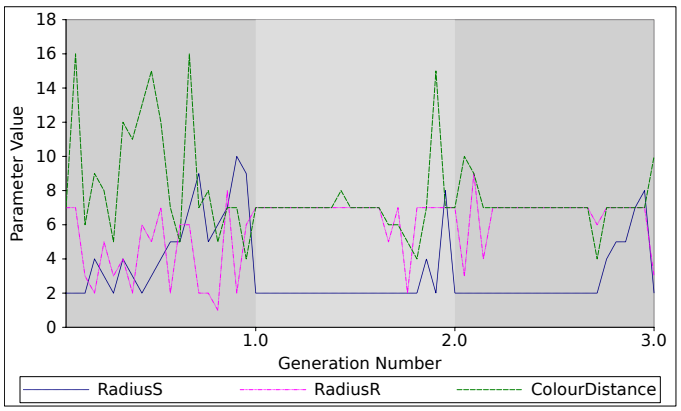

Figure 5. The first three generations of the meanshift GA evaluation