A Supervised Approach to the Evaluation of Image Segmentation Methods

Luren Yang, Fritz Albregtsen, Tor Lønnestad and Per Grøttum

Department of Informatics, University of Oslo
P.O. Box 1080, Blindern, N-0316 Oslo, Norway

Abstract. Evaluation is an important step in developing a segmentation algorithm for an image analysis system. We first give a review of segmentation evaluation methods, and then demonstrate how a supervised evaluation method based on shape features is used in the development of a segmentation algorithm for fluorescence images of white blood cells.

1 Introduction

An image analysis system typically has three major components: image segmentation, feature extraction, and feature analysis. Segmentation is to divide an image into meaningful regions such as object and background. Features of each region are then estimated for further analysis. The accuracy of segmentation is often crucial to the performance of an image analysis system. Many segmentation methods have been proposed [3]. The selection of a proper segmentation method for a particular problem is often based on testing and evaluation. Therefore, the evaluation of image segmentation methods is an interesting research topic, which has already been discussed by some authors [17, 12, 6, 7, 18, 19, 11].

We divide the evaluation methods into two groups: supervised and unsupervised evaluation, depending on whether the method utilizes a priori knowledge of a reference segmentation. In supervised evaluation, the difference between a reference segmentation and the output of a segmentation algorithm is computed. We demonstrate how the method is used in the development of a segmentation algorithm for fluorescence images of white blood cells. In our example, three manual segmentations are combined as a reference.

2 A Review of Evaluation Methods

Unsupervised evaluation does not depend on a correct segmentation. Haralick and Shapiro [3] established some qualitative guidelines for a good image segmentation. Quantitative segmentation performance measures were developed by several authors for unsupervised evaluation. Weszka and Rosenfeld [12] used a busyness measure and a classification error as the performance criteria. Levine and Nazif [7] defined a set of parameters for unsupervised evaluation including region uniformity, region contrast, line contrast and line connectivity. They assumed that some features should be uniform inside each region, and distinct between adjacent regions. Sahoo *et al.* [10] used the uniformity criterion of Levine and Nazif [7] and a shape measure, computed from the gradient values and the selected threshold value. The local thresholding method of Yanowitz and Bruckstein [16] assumes that the boundary of an object is located at the edge where the magnitude of intensity gradient is large. Therefore the average gradient magnitude along the boundary is used as a criterion in their "validation step". We note that the average gradient magnitude can be considered as an unsupervised performance measure.

A supervised evaluation utilizes a reference segmentation and measures the difference between the reference segmentation and the output of a segmentation algorithm. The simplest supervised measure is the probability of error. For an object-background image, it is defined as

$$P(\text{error}) = P(O)P(B|O) + P(B)P(O|B) \tag{1}$$

where $P(R_1|R_2)$ is the probability of classifying R_2 as R_1, and $P(R_1)$ is the a priori probability of class R_1. This measure is often used in the evaluation of thresholding techniques (for example in [5, 1]). Albregtsen [1] computed the probability of error as a function of the object-to-background area ratio, and evaluated the performance of thresholding methods working at a low object-to-background ratio. The probability of error may also be a criterion for developing an optimal thresholding method. The probability of error for images with many classes was formulated and applied by Lim and Lee [8].

As already pointed out by Yasnoff *et al.* [17], the probability of error may not give sufficient information about the error. Yasnoff *et al.* [17] proposed to use, in addition to the probability of error, another measure called pixel distance error in which the positions of the misclassified pixels were taken into account. Levine and Nazif [6] proposed to separate the probability of error into an under-merging error and an over-merging error. In many image analysis tasks, the ultimate goal of segmentation is to obtain measurements of the object features. Let x be a feature computed from an object in a reference image, and $\hat{x}$ be that computed from the object in the output of a segmentation algorithm. The ultimate measurement accuracy (UMA) proposed by Zhang and Gerbrands [18] is defined by

$$\text{UMA} = |x - \hat{x}|. \tag{2}$$

Trier and Jain [11] developed a goal-directed method to evaluate the segmentation of document images with the purpose of symbol recognition. This method uses three quantitative measures which are computed from the result of symbol recognition.

3 A Supervised Evaluation

We illustrate our evaluation strategy through an example. Before giving the evaluation method, we briefly present the segmentation algorithm and discuss some shape features to be used in the evaluation process.

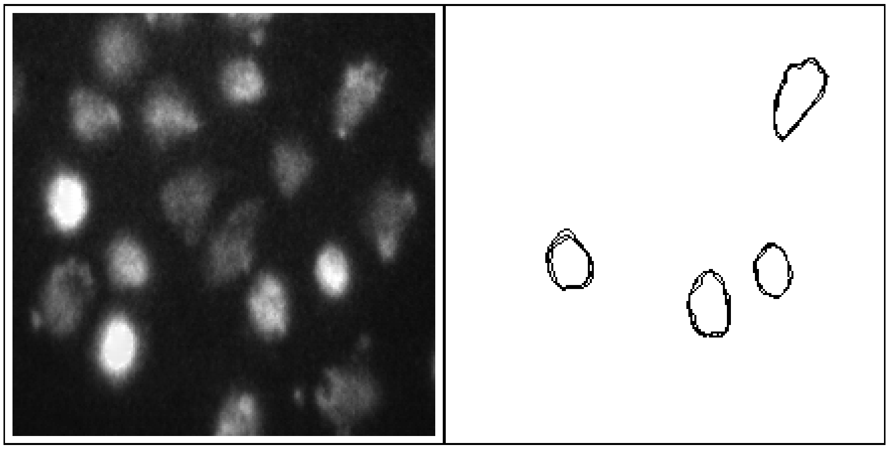

Fig. 1. (left) A gray level fluorescence cell image. (right) Boundaries of 4 cells in the image to the left, marked by three different persons.

3.1 A Segmentation Algorithm

We have developed an image analysis system for measuring shape and motion of white blood cells from a sequence of fluorescence microscopy images [15]. The purpose is to study the relation between the intracellular calcium concentration and the cell motion. We first segment the image. Shape features are then computed from each cell object, and used to describe cell shape and motion. Figure 1 shows a gray level cell image together with the result of a manual segmentation of 4 cells done by three physiologists.

We have used a two-pass automatic segmentation algorithm [15]. In the first pass, an initial segmentation is applied to classify the pixels into cell and background. Region labelling, correction and cell tracking are then done in the second pass to give a final segmentation. Special methods are developed for the second pass. In the first pass we have tried many standard segmentation methods. Global thresholding methods do not work well since the images have uneven background. The Laplace of Gaussian (LoG) method, the local thresholding method of Eikvil *et al.* [2] and a dynamic thresholding method which we call the modified Bernsen's method [15] give good results according to visual examination. These three methods are then compared by a quantitative evaluation. As a preprocess to the segmentation, we smooth the images by a Gaussian filter whose parameter is also determined according to the quantitative evaluation.

3.2 Shape Features

We use two-dimensional shape features to quantify the difference between two segmentations. The area A and the perimeter P of an object are two features related to the size of the object. The circularity C defined by $C = 4\pi A/P^2$ is a scaling invariant shape feature. Kulpa's method [4] has been used to compute the cell perimeter. (See [14] for an evaluation of several area and perimeter estimators.) Cartesian geometric moments of binary regions have been efficiently computed [13] in order to obtain moment-based shape features [9] such as area, centroid, radius of gyration, orientation, and image ellipse. The elongation has been measured as the ratio of the lengths of the semimajor and semiminor axes of the image ellipse.

3.3 The Evaluation Method

The manual segmentation as shown in Figure 1(right) is used as the reference segmentation. To measure the difference between the reference and the output of a segmentation algorithm, we first compute an under-merging error (UM) and an over-merging error (OM). These two measures were first proposed by Levine and Nazif [6]. We have modified them so that they are computed for each object, and the size of the object is taken into consideration. We assume that a reference image has three regions: a background region (B) consisting of the pixels classified as background by all three manual segmentations, an object region (O) consisting of the pixels classified as object by all the manual segmentations, and an uncertain region consisting of all the other pixels. The output of the automatic segmentation algorithm has two regions: a background ($\hat{B}$) and an object ($\hat{O}$) region. The under-merging error (UM) and the over-merging error (OM) are then defined by

$$\mathrm{UM} = \mathrm{area}(O \backslash \hat{O})/A \qquad \mathrm{OM} = \mathrm{area}(B \backslash \hat{B})/A \tag{3}$$

where the difference operation $R_1 \backslash R_2$ is defined by

$$R_1 \backslash R_2 = \{p | p \in R_1,\, p \notin R_2\} \tag{4}$$

in which area(R) is the area of region R, and A is the average of the object areas obtained by the three manual segmentations. A total difference measure (DM) is then computed as

$$\mathrm{DM} = \mathrm{UM} + \mathrm{OM} \tag{5}$$

To obtain more information about the difference, we use the UMA of Zhang and Gerbrands [18] for some selected shape features. We compute the UMA for circularity and orientation using equation (2). Since there are three reference images for each cell, the correct feature value x is computed by averaging the three. For area and elongation, we normalize the UMA by the average of the feature values estimated from the three manual segmentation results, i.e.

$$\mathrm{UMA} = |x - \hat{x}|/x \tag{6}$$

The values of the difference measures (UM, OM and DM) and the UMA values are first computed for each object, and then averaged over all the objects. Their standard deviations are also computed, because a systematic error with small standard deviation should be considered less harmful than an error with large standard deviation.

Table 1. The means and standard deviations of the measures of differences between the reference segmentation and the output of the automatic segmentation algorithm.

Segment.	Gauss.	UM		OM		DM	
Method	σ	mean	std.	mean	std.	mean	std.
LoG	2.0	0.027	0.025	0.091	0.126	0.119	0.119
	3.0	0.019	0.022	0.060	0.072	**0.079**	**0.071**
	4.0	0.008	0.015	0.102	0.087	0.115	0.084
Eikvil	1.0	0.068	0.060	0.009	0.020	0.077	0.056
	2.0	0.051	0.047	0.018	0.032	**0.070**	**0.046**
	3.0	0.039	0.038	0.039	0.055	0.078	0.056
	4.0	0.029	0.030	0.057	0.079	0.086	0.077
Bernsen	1.0	0.045	0.039	0.018	0.020	0.063	0.050
	2.0	0.024	0.035	0.038	0.039	**0.063**	**0.041**
	3.0	0.017	0.027	0.082	0.076	0.100	0.073
	4.0	0.010	0.020	0.101	0.086	0.111	0.080

Table 2. The means and standard deviations of the UMA values. UMA $= |x - \bar{x}|/x$ for area and elongation, and UMA $= |x - \bar{x}|$ for circularity and orientation (in degrees).

Segment.	Gauss.	Area		Circularity		Orientation		Elongation	
Method	σ	mean	std.	mean	std.	mean	std.	mean	std.
LoG	2.0	0.189	0.209	0.144	0.114	8.804	13.966	0.267	0.245
	3.0	**0.109**	0.103	0.055	0.047	4.627	5.151	0.198	0.195
	4.0	0.182	0.122	**0.041**	**0.033**	2.467	1.938	0.154	0.158
Eikvil	1.0	0.173	0.113	0.071	0.056	6.348	8.284	0.127	0.131
	2.0	0.144	0.092	0.084	0.059	2.205	1.810	0.138	0.130
	3.0	0.122	0.098	0.107	0.058	6.658	6.903	0.162	0.142
	4.0	0.169	0.120	0.122	0.059	7.854	7.617	0.192	0.161
Bernsen	1.0	0.120	0.096	0.065	0.038	5.079	5.671	0.108	0.107
	2.0	0.112	**0.086**	0.079	0.055	**2.097**	**1.614**	**0.075**	**0.059**
	3.0	0.154	0.111	0.106	0.057	6.165	6.287	0.162	0.137
	4.0	0.169	0.120	0.120	0.058	7.131	7.254	0.191	0.160
Manual		0.081	0.064	0.031	0.027	1.907	1.709	0.078	0.070

3.4 Results

We evaluated three segmentation methods (the modified Bernsen's method, the LoG method, and the method of Eikvil *et al.*) combined with Gaussian smoothing with different filter standard deviations (σ = 1.0, 2.0, 3.0 and 4.0). The LoG method was very sensitive to the noise and broke down when $\sigma = 1.0$.

The means and the standard deviations of the UM, OM and DM values computed from 100 cell objects are shown in Table 1. The smallest mean and standard deviation of the DM values of each segmentation algorithm are indicated in boldface. Table 2 shows the means and the standard deviations of the UMA values for area, circularity, orientation and elongation. In each column of the table, the smallest value obtained through an automatic segmentation is indicated in boldface. The UMA values were also computed for the manual segmentation and given in the last row of the table to show how a feature obtained from a manual segmentation differs from the average of the three.

The results show that the modified Bernsen's method with $\sigma = 2.0$ is the best. It gives the least errors for many of the measures listed in the two tables. With this method, the UMA means and standard deviations are not much larger than the values from the manual segmentation. We can see that the smoothing filter has a clear effect on the result, making the objects larger. When the degree of blur increases, the OM increases and the UM decreases. Also, different segmentation methods may require different degrees of smoothing.

4 Discussion and Conclusion

As we show in this paper, the segmentation process itself can be done step by step, and the algorithm of each step can be chosen by an evaluation. Basically, there are two types of evaluation: supervised and unsupervised evaluation.

An unsupervised evaluation often favors a segmentation which gives uniform and homogeneous regions. However, such segmentation does not necessarily give accurate boundaries between different regions.

The criteria of a supervised evaluation is the difference between the output of a segmentation algorithm and a reference segmentation. If the segmentation algorithm is applied to a synthetic test image, one will have a correct segmentation as the ground truth. However, for many image analysis tasks, it is not practical to make synthetic test images since the real world is not easily modelled. One common method is then to make the reference by a manual segmentation, since a human being is often the best judge of the machine output. To reduce the random error made by a human being, we suggest to make several manual segmentations for one image, probably by different persons. We have shown that several manual segmentations can be combined as a reference.

When the reference is available, the evaluation is to quantify the difference between the reference and the machine output. In many cases, one performance measure is not enough to describe the difference between the two. The total difference measure (DM) may be used to assess the magnitude of the difference, and the other measures (OM, UM, and UMA values) are available to describe the nature of the difference. We show how the evaluation method is used in the development of a particular segmentation algorithm for fluorescence white blood cell images. The purpose of this evaluation is to choose an appropriate segmentation method among three, and to decide the proper value of a parameter of a smoothing filter. One segmentation method (the modified Bernsen's method) is chosen since the means and the standard deviations of the errors are small. The error standard deviation shows whither the error is random or systematic. We also measure the difference between the manual segmentation results, in order to compare the error made by machine and the random error made by a human being. This comparison indicates the reliability of a machine segmentation algorithm.

References

1. Albregtsen, F.: Non-parametric histogram thresholding methods – error versus relative object area. Proc. 8th Scadinavian Conf. Image Analysis (1993) 273–280
2. Eikvil, L., Taxt, T., Moen, K.: A fast adaptive method for binarization of documents images. Proc. 1st Int. Conf. Document Analysis and Recognition (1991)
3. Haralick, R. M., Shapiro, L. G.: Image segmentation techniques. Comput. Vision Graph. Image Process. **29** (1985) 100–132
4. Kulpa, Z.: Area and perimeter measurement of blobs in discrete binary pictures. omput. Graph. Image Process. **6** (1977) 434–451
5. Lee, S. U., Chung, S. Y., Park, R. H.: A comparative performance study of several global thresholding techniques for segmentation. Comput. Vision Graph. Image Process. **52** (1992) 171–190
6. Levine, M. D., Nazif, A. M.: An experimental rule-based system for testing low level segmentation strategies. Multicomputers and Image Processing Algorithms and Programs. Academic Press (1982) 149–160
7. Levine, M. D., Nazif, A. M.: Dynamic measurement of computer generated image segmentations. IEEE Trans. Pattern Anal. Machine Intell. **7** (1985) 155–164
8. Lim, Y. W., Lee, S. U.: On the color image segmentation algorithm based on the thresholding and the fuzzy c-means techniques. Patt. Recogn. **23** (1990) 935–952
9. Prokop, R. J., Reeves, A. P.: A survey of moment-based techniques for unoccluded object representation and recognition. CVGIP: Graphical Models and Image Processing **54** (1992) 438–460
10. Sahoo, P. K., Soltani, S., Wong, A. K. C., Chen, Y. C.: A survey of thresholding techniques. Comput. Vision Graph. Image Process. **41** (1988) 233–260
11. Trier, Ø. D., Jain, A. K.: Goal-directed evaluation of binarization methods. Proc. NSF/ARPA Workshop on Performance vs. Methodology in Computer Vision (1994)
12. Weszka, J. S., Rosenfeld, A.: Threshold evaluation techniques. IEEE Trans. Sys. Man Cyb. **8** (1978) 622–629
13. Yang, L., Albregtsen, F.: Fast computation of invariant geometric moments: a new method giving correct results. Proc. 12th Int. Conf. Pattern Recognition, Vol. I (1994) 201–204
14. Yang, L., Albregtsen, F., Lønnestad, T., Grøttum, P.: Methods to estimate areas and perimeters of blob-like objects: a comparison. Proc. IAPR Workshop on Machine Vision Applications (1994) 272–276
15. Yang, L., Albregtsen, F., Lønnestad, T., Grøttum, P., Iversen, J.-G., Røtnes, J. S., Røttingen, J.-A.: Measuring shape and motion of white blood cells from sequences of fluorescence microscopy images, Proc. 9th Scandinavian Conf. Image Analysis, Vol. I (1995) 219–227
16. Yanowitz, S. D., Bruckstein, A. M.: A new method for image segmentation. Comput. Vision Graph. Image Process. **46** (1989) 82–95
17. Yasnoff, W. A., Mui, J. K., Bacus, J. W.: Error measures for scence segmentation. Pattern Recognition **9** (1977) 217–231
18. Zhang, Y. J., Gerbrands, J. J.: Segmentation evaluation using ultimate measurement accuracy. Proc. SPIE Vol. 1657, Image Processing Algorithms and Techniques III (1992) 449–460
19. Zhang, Y. J.: Segmentation evaluation and comparison: a study of various algorithms. Proc. SPIE Vol. 2094, Visual Communications and Image Processing (1993) 801–812