ON SUPERVISED METRICS FOR SHAPE SEGMENTATION*

Dibet Garcia Gonzalez
Advanced Computer Sciences Technologies Center, Ciego de Avila University, Moron Street, Ciego de Avila, Cuba

Miguel Garcia Silvente
Computer Science and Artificial Intelligence Department, Granada University, Granada, Spain

Keywords: Segmentation evaluation, Metrics, Comparison functions, Segmentation algorithms, Thresholding, Ranking.

Abstract: Segmentation is one of the most critical steps in image analysis. Also, the quantification of the error commited during this step is not a straightforward task. In this work, the performance of some comparison function or metrics are studied, when just one object appears in the analyzed regions. We develop a method for rank many validation measures of segmentation algorithms. It is based on thresholding a test image with a range of threshold and to find the middle threshold value when the performance measure is minimum or maximum. The performance is plotted and the first derivate is employed in the ranking construction. We have determined that RDE and MHD are two performance measures that show the best results (both are the most selective).

1 INTRODUCTION

One of the fields in which segmentation is a critical task is biomedicine. Many medical tests are based on the study of medical images. Some examples are the detection of breast cancer (Kiyan and Yildirim, 2004; Joo et al., 2004), uterine cervix cancer (Yang-Mao et al., 2008) and the morfologic study of brain sub structures like the hypocampus and the amygdalas (Shentona et al., 2002). A medical specialist can make a mistake when processing many images in a short period of time. They can arrive to different conclusions in two different moments with the same image (intra observator errors), and also two or more medical specialists can arrive to different conclusions with the same picture (inter observator errors). The knowledge of the specialist has an important role in this process. The computer vision field is very important in the automatization of many processes that are tedious for the human being.

The image segmentation[2] is an important step in the computer vision system. Its effectivity is influenced by the quality of the segmentation previously done. For that reason, the development of new segmentation algorithms has the attention of many specialists. Watershed (Meyer, 1992), Snakes (Kass et al., 1988), Region Growing (Adams and Bischof, 1994) and Mean Shift (Comaniciu and Meer, 1999) are some segmentation algorithms developed for general purpose. The evaluation of those algorithms has a crucial importance due to the ethical issues associated with a wrong diagnosis.

In the evaluation process, the selection of a comparison function or performance measure is also very important. Some authors employ some of them when other authors employ others or a set of them. It does not exist a valoration about the performance of each published measure. This work pays attention on this issue.

It has been divided in four sections. Sec. 2 includes some comparison functions related to the literature, Sec. 3 points out some important considerations for the analysis of the measures and it presents a ranking method for segmentation metrics. Finally, Sec. 4 shows experiments and results and Sec. 5 shows the conclusions and directions for future works.

*This work is supported by Spanish MCI Project TIN2007-66367 and Andalusian Regional Government project TIC1670.

[2]A segmentation algorithm can extract one or more regions of interest in an image.

Garcia Gonzalez D. and Garcia Silvente M. (2010).
ON SUPERVISED METRICS FOR SHAPE SEGMENTATION.
In *Proceedings of the Third International Conference on Bio-inspired Systems and Signal Processing*, pages 468-473
DOI: 10.5220/0002759004680473

2 THE COMPARISON FUNCTIONS

In the literature some comparison function appears. They are classified in supervised and unsupervised depending if it needs a reference image or ground truth (GT). The unsupervised function does not need a GT while the supervised does. A recent survey of unsupervised methods can be studied in (Zhang et al., 2008). In addition, some supervised measures can be found in the folowing works (Cavallaro et al., 2002; Janasievicz et al., 2005; Baddeley, 1992; Dice, 1945; Sezgin and Sankur, 2004; Cardoso and Corte-Real, 2005; Dubuisson and Jain, 1994; Ge et al., 2006; Pratt et al., 1978; Martin et al., 2001; Polak M, 2009; Pratt, 1997; Popovic et al., 2007; Yang-Mao et al., 2008; Monteiro and Campilho, 2006; Boucheron et al., 2007). This work is related to supervised measures.

The segmentation results (SR) and the GT can contain one or more objects in an image. They are many metrics that can just evaluate one SR object versus one GT object. Some examples are relative distance error (RDE) (Yang-Mao et al., 2008), the Hausdorff distance (HD) and the modified Hausdorff distance (MHD) (Dubuisson and Jain, 1994), coverage factor (CF) (Popovic et al., 2007), dice coeficient (DSC) (Dice, 1945), relative area error (RAE) (Sezgin and Sankur, 2004; Yang-Mao et al., 2008), P measure (PM) (Ge et al., 2006) and figure of merit (FOM) (Pratt et al., 1978).

Furthermore, a multiobject performance measure can be simplified for only one object. In this work all the multi object performance measures presented are simplified in order to evaluate the results of only one object by image. For instance, global consistency error (GCE) (Martin et al., 2001), local consistency error (LCE) (Martin et al., 2001), misclasification error (ME) (Sezgin and Sankur, 2004) and object-level consistency error (OCE) (Polak M, 2009).

All the measures mentioned previosly can be found in Eq. 1, Eq. 2, Eq. 3, Eq. 4, Eq. 5, Eq. 6, Eq. 7, Eq. 8, Eq. 9 and Eq. 10.

If $e_1, e_2, e_3, \ldots, e_{n_e}$ are the SR pixels and $t_1, t_2, t_3, \ldots, t_{n_t}$ are the GT pixels, n_e and n_t are the number of pixels of SR and GT, respectively; $dist(e_i, t_j)$ represents an euclidean distance between e_i and t_j; $d_{t_j} = min\{dist(e_i, t_j) \,|\, i = 1, 2, \ldots, n_e\}$ and $d_{e_i} = min\{dist(e_i, t_j) \,|\, j = 1, 2, \ldots, n_t\}$ then:

$$RDE = \frac{1}{2}\left(\sqrt{\frac{1}{n_e}\sum_{i=1}^{n_e} d_{e_i}^2} + \sqrt{\frac{1}{n_t}\sum_{j=1}^{n_t} d_{t_j}^2}\right) \quad (1)$$

$$MHD = max\{mean_{e_i}\{d_{e_i}\}, mean_{t_j}\{d_{t_j}\}\} \quad (2)$$

$$HD = max\{max_{e_i}\{d_{e_i}\}, max_{t_j}\{d_{t_j}\}\} \quad (3)$$

If TP, FP, FN, TN are elements of the confusion matrix; p is the sensibility and q the specificity calculated from the confusion matrix; A is the area of the GT object and B is the area of the SR object then:

$$GCE = LCE = min\left(\left(1 - \frac{TP}{A}\right), \left(1 - \frac{TP}{B}\right)\right) \quad (4)$$

$$RAE = \left\{ \begin{array}{ll} \frac{FP-FN}{TP+FP}, & FP \geq FN; \\ \frac{FN-FP}{TP+FN}, & FP < FN \end{array} \right\} \quad (5)$$

$$CF = \left\{ \begin{array}{ll} d, & p > q \wedge p > 1-q; \\ -d, & p < q \wedge p > 1-q; \\ undefined, & p \leq 1-q. \end{array} \right\} \quad (6)$$

where $d = \frac{2p(1-q)}{p+(1-q)} + \frac{2(1-p)q}{(1-p)+q}$

$$1 - DSC = 1 - \frac{2TP}{2TP+FP+FN} \quad (7)$$

$$ME = 1 - \frac{TN+TP}{TN+FN+FP+TP} \quad (8)$$

$$OCE = 1 - PM = 1 - \frac{TP}{TP+FP+FN} \quad (9)$$

$$1 - FOM = 1 - \frac{1}{\mu}\sum_{l=1}^{\mu}\left(\frac{1}{1+k\cdot d_{e_i}^2}\right) \quad (10)$$

where $\mu = max\{n_e, n_t\}$ and k is a scale factor.

All the measures are dissimilarity functions except PM, DSC and FOM whichs are converted substracting one minus the measure value. This is important because it will determine if the middle threshold value will be reached for a minimum or for a maximum. RDE, MHD and HD are not normalizated measures. In the Fig. 2, for ploting propose, they are normalizated dividing every element value by the maximum value of the measure.

3 SOME CONSIDERATIONS ABOUT THE COMPARISON FUNCTIONS

Many authors think that only in the context of the full system evaluation, the efectiveness of a segmentation algorithm can be determined (Everingham et al., 2001). The use of comparison functions, in low level evaluation, is getting more attention nowadays. The supervised or unsupervised measures are employed individually or combined. Either as the average

of some functions (Yang-Mao et al., 2008), or a fitness/cost function (Everingham et al., 2001), or using machine learning (Zhang et al., 2006).

Some authors (Vilalta and Oblinger, 2000; Huang and Ling, 2005; Popovic et al., 2007) employ the statistical criteria of consistency and discriminancy in the evaluation of the performance measures. They allow to determine if two performance measures are consistent among them and which is more discriminant. If two performance measures are consistent, the best metric is the most discriminant. By other side, the proposed measure in (Janasievicz et al., 2005) is compared with others by evaluating the resistance of segmentation methods to noise, shrinking and stretching. That method evaluates the sensitivity of the performance measure. Rosenberger (Rosenberger, 2006) shows a psychovisual study made with 160 experts. They rank the segmentation results of some images. Then, he compares seven supervised measures results with the experts evaluation allowing to determinate which is the best choice according to the human judgement. Those criteria are not the objective of this paper. They will be analysed in a future work.

In (Freixenet et al., 2002) is presented a comparative survey of seven of the most frequently used strategies to perform segmentation based on region and boundary information. Concretely, on the one hand, algotihms based on the region evaluation parameters produce the best results. And on the other hand, algorithms based on the boundary evaluation parameters are the best ones. These discrepancies stimulate the develop of a ranking method to compare functions of segmentation algorithms.

In (Zhang, 1996) is shown a method for ranking many validation metrics. It allows to select the best performance measure. It consists in thresholding an image for a threshold range [low..high] and then plot all the comparison functions results. All the measures should have a middle value where the performance measure is minimum. "Comparing the depth of valleys, these methods can be ranked in order" (Zhang, 1996). The deepest valley corresponds to the best performance measure. We do not agree with that assumption because many measures are not normalized between zero and one. Besides, they employ different theories (distances, areas, confusion matrix and so on) in the calculation. It can be deduced from the (Zhang, 1996) proposition that most of the measures should match with the same threshold value when it's reached its maximum or minimum. The minimum or maximum is determinate by the use of a dissimilarity or similarity comparison function, respectively.

The proposed method consists of doing a better analysis from the study of the first derivative of the curves previously refered, the curve that its first derivative increases quicker and decreases faster than the rest. It means that the measures respond faster to variations in the quality of the segmentation result. That characteristic makes a ranking study of the first derivative of the curves.

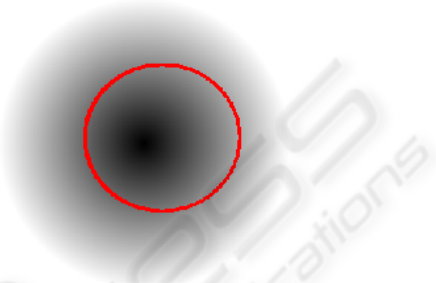

Figure 1: Test image for thresholding. The red circle corresponds to the GT.

4 EXPERIMENTS AND RESULTS

In this section we present an example in order to show the proposed method and some special cases.

4.1 Experiments

Some questions are taken into account to ease the analysis but without removing the rigor of this work.

- The performance measures RDE, MHD and HD are normalizated dividing by its maximum value because they are not in the range [0..1]. That is only for graphic purpose and it is plotted in Fig. 2. In Fig. 3 they appear not normalizated.
- The absolute value of CF (ABS(CF)) is considered because it is the only measure in the range [-1..1].

The experiment consists in thresholding the image in Fig. 1 (without the red circle) between a low value and a high value ([75..139] in this case) and ploting the performance using all the comparison functions shown in Sec 2. When the threshold is increasing from low to high, the performance should be better until some middle threshold is reached. From the middle to high the performance should increase again (Zhang, 1996).

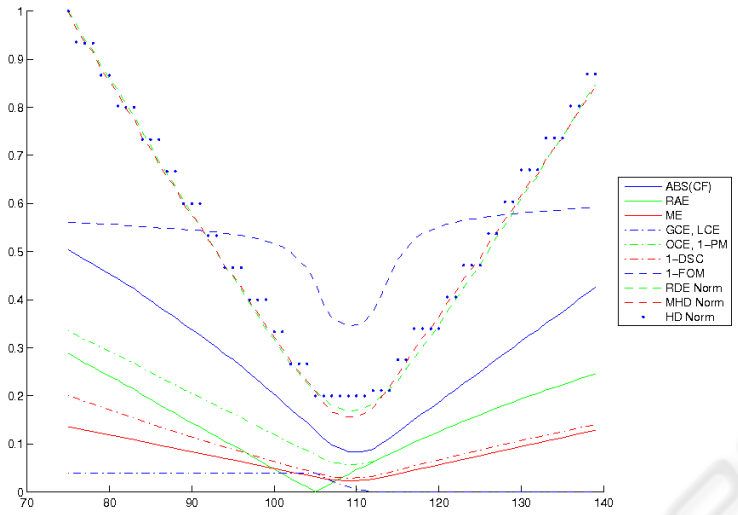

Figure 2: Results of thresholding the picture in Fig. 1, when the red circle is not present, as the GT.

4.2 Results

Fig. 2 shows the results of the experiment previously mentioned.

From Fig. 2 can be found that GCE and LCE have the worst performance because they have low variation in all the range examined. Only in a short range [104..113] the performance measure shows visible variations. For the range [113..139] the performance measure value equals to zero means the SR and the GT are equal, being this not right. This experiment shows that GCE and LCE could not be employed to evaluate segmentation algorithms in which GT and SR have only one object. GCE is equal to LCE when they are simplified.

1-FOM is not a good performance measure because, although in the range [99..119] the performance measure has an expected performance, in the range [75..99] and the range [119..139] it has low variations ocurring a similar performance as GCE and LCE. Another reason is the existence of a parameter to adjust, which is not covenient.

RAE is a metric that shows an apparently good performance but when they are compared with the rest, the valley of its curve is reached for the threshold 105 while for all the rest, it is reached in 110. It means that is inconsistent with the rest.

The Hausdorff distance has many irregularities in its curve therefore this performance measure is not convenient for the segmentation algorithms evaluation. These irregularities appears because the performance measure shows the same result for different SR.

The other metrics (RDE, MHD, 1-DSC, OCE, 1-PM, ME, CF) have a good performance in the selected range. Coverage Factor has the main advantage that is the only measure that can distinguish between over and under segmentation. Next, the last group of metric are analysed more precisely.

Another kind of result can be shown more clearly observing Fig. 3. It shows the first derivative of the curves RDE, MHD, 1-DSC, OCE, 1-PM, ME, CF shown in Fig. 2. Now, we propose to make a ranking observing which performance measure grows and decreases faster. The examination of Fig. 3 allows to determine the following ranking (from more selective to less selective):

1. RDE, 2. MHD, 3. CF, 4. CCE, 1-PM, 5.DSC, 6. ME

The other measure shows an undesireable performance as it was previously demonstrated.

5 CONCLUSIONS AND FUTURE WORK

The evaluation of the segmentation is considered a hard task in the field of computer vision. As the same number of specialists can emit different views about the quality of the results of several segmentation algorithms, it happens that several measures may differ in selecting the best segmentation algorithm.

In this work, a ranking method to compare functions is presented. It permits to analyze the selec-

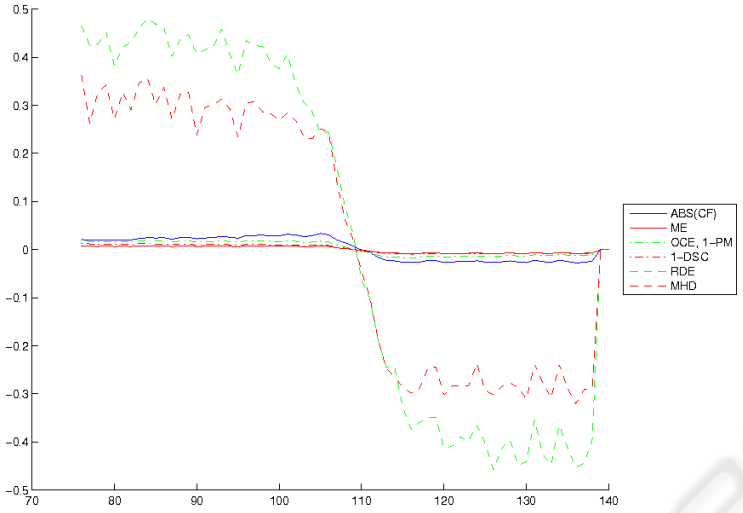

Figure 3: First derivative of the curves RDE, MHD, 1-DSC, OCE, 1-PM, ME, CF shown in Fig. 2.

tivity of the measure. A good selectivity is a desired property for a comparison function. The metrics are used in the evaluation of a thresholded image within a threshold range. That allows to study the performance of the comparison functions. So, the results and its first derivative are plotted in two graphics. The first graphic shows if the performance measure is able to find the same best threshold value. The second one shows which comparison function grows and decreases faster and so it provides a ranking. We conclude that RDE and MHD are two performance measures that show the best results and are more selective than the rest ones. Other metrics like GCE and LCE are not recomended for one GT object evaluation.

We plan to use the criteria of consistency and discrimination to evaluate the behavior of the comparison functions. Another possibility is to extend the measure RDE to include special cases which were not initially taken into account.

ACKNOWLEDGEMENTS

We would like to thank to the MAEC-AECID for the grant to the first author.

REFERENCES

Adams, R. and Bischof, L. (1994). Seeded region growing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16(6):641–647.

Baddeley, A. (1992). Errors in binary images and an L^p version of the hausdorff metric. *Nieuw Archief voor Wiskunde*, 10:157–183.

Boucheron, L., Harvey, N., and Manjunath, B. (2007). A quantitative object-level metric for segmentation performance and its application to cell nuclei. *Lecture Notes of Computer Science*, 4841:208–219.

Cardoso, J. S. and Corte-Real, L. (2005). Toward a generic evaluation of image segmentation. *IEEE Transactions on Image Processing*, 14(11):1773–1782.

Cavallaro, A., Gelasca, E. D., and Ebrahimi, T. (2002). Objetive evaluation of segmentation quality using spatio-temporal context. In *IEEE International Conference on Image Processing*, Rochester (New York).

Comaniciu, D. and Meer, P. (1999). Mean shift analysis and applications. *IEEE International Conference on Computer Vision*, 2:1197.

Dice, L. R. (1945). Measures of the amount of ecologic association between species. *Ecology*, 26:297–302.

Dubuisson, M. P. and Jain, A. K. (1994). A modified hausdorff distance for object matching. In *Proceedings of 12th International Conference on Pattern Recognition*, pages A–566–A–569.

Everingham, M., Muller, H., and Thomas, B. (2001). Evaluating image segmentation algorithms using monotonic hulls in fitness/cost space. In *Proceedings of the British Machine Vision Conference*, pages 363–372.

Freixenet, J., Muoz, X., Raba, D., Mart, J., and Cuf, X. (2002). Yet another survey on image segmentation: Region and boundary information integration. *Lecture Notes of Computer Science*, 2352:408–422.

Ge, F., Wang, S., and Liu, T. (2006). Image-segmentation evaluation from the perspective of salient object extraction. In *Proceedings of the 2006 IEEE Computer*

Society Conference on Computer Vision and Pattern Recognition.

Huang, J. and Ling, C. X. (2005). Using auc and accuracy in evaluating learning algorithms. *IEEE Transactions on Knowledge Data Eng*, 17(3):299–310.

Janasievicz, R., Pinheiro, G., and Facon, J. (2005). Measuring the quality evaluation for image segmentation. *Lecture Notes of Computer Science*, 3773:120–127.

Joo, S., Yang, Y., Moon, W., and Kim, H. (2004). Computer-aided diagnosis of solid breast nodules: Use of an artificial neural network based on multiple sonographic features. *IEEE Transactions on Medical Imaging*, pages 1292–1300.

Kass, M., Witkin, A., and Terzopoulos, D. (1988). Snakes: Active contour models. *International Journal of Computer Vision*, pages 321–331.

Kiyan, T. and Yildirim, T. (2004). Breast cancer diagnosis using statistical neural networks. *Journal of Electrical & Electronics Engineering*, 4:1149–1153.

Martin, D., Fowlkes, C., Tal, D., and Malik, J. (2001). A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *Proc. 8th International Conference Computer Vision*, volume 2, pages 416–423.

Meyer, F. (1992). Color image segmentation. In *Proceedings of the International Conference on Image Processing and Its Application*, pages 303–306.

Monteiro, F. and Campilho, A. (2006). Performance evaluation of image segmentation. *Lecture Notes of Computer Science*, 4141:248–259.

Polak M, Zhang H, P. M. (2009). An evaluation metric for image segmentation of multiple objects. *Image and Vision Computing*, 27(8):1223–1227.

Popovic, A., de la Fuente, M., Engelhardt, M., and Radermacher, K. (2007). Statistical validation metric for accuracy assessment in medical image segmentation. *International Journal of Computer Assisted Radiology and Surgery*, 2:169–181.

Pratt, W. K. (1997). *Digital image processing*. John Wiley and Sons, New York.

Pratt, W. K., Faugeras, O. D., and Gagalowicz, A. (1978). Visual discrimination of stochastic texture fields. *IEEE Transactions on Systems, Man and Cybernetics Part B*, 8(11):796–804.

Rosenberger, C. (2006). Adaptative evaluation of image segmentation results. *International Conference on Pattern Recognition*, 2:399–402.

Sezgin, M. and Sankur, B. (2004). Survey over image thresholding techniques and quantitative performance evaluation. *Journal of Electronic Imaging*, 13(1):146–165.

Shentona, M., Gerigb, G., McCarleya, R., Szekelyc, G., and Kikinisd, R. (2002). Amygdalahippocampal shape differences in schizophrenia: the application of 3d shape models to volumetric mr data. *Psychiatry Research Neuroimaging*, 115:15–35.

Vilalta, R. and Oblinger, D. (2000). A quantification of distance bias between evaluation metrics in classification. In *Proceedings of the Seventeenth International Conference on Machine Learning*, pages 1087–1094.

Yang-Mao, S.-F., Chan, Y.-K., and Chu, Y.-P. (2008). Edge enhancement nucleus and cytoplast contour detector of cervical smear images. *IEEE Transactions on Systems, Man and Cybernetics Part B: Cybernetics*, 38(2):353–366.

Zhang, H., Cholleti, S., Goldman, S. A., and Fritts, J. (2006). Meta-evaluation of image segmentation using machine learning. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*.

Zhang, H., Frits, J. E., and Goldman, S. A. (2008). Image segmentation evaluation: A survey of unsupervised methods. *Computer Vision and Image Understanding*, 11:260–280.

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