A COMPARATIVE STUDY OF SUPERVISED EVALUATION CRITERIA FOR IMAGE SEGMENTATION.

S. Chabrier, H. Laurent, B. Emile, C. Rosenberger and P. Marché

Laboratoire Vision et Robotique, UPRES EA 2078
ENSI de Bourges / Université d'Orléans
10 boulevard Lahitolle, 18020 Bourges Cedex, France (Europe)
phone: +33 2 48484000, fax: +33 2 48484040, email: sebastien.chabrier@ensi-bourges .fr

ABSTRACT

This paper presents a comparative study of five supervised evaluation criteria for image segmentation. The different criteria have been tested on a selection of hundred images extracted from the ©Corel database for which manual segmentation results provided by experts are available. Nine segmentation algorithms have been considered, most of which are based on threshold selection. In order to compare the behavior of the different criteria towards over- and undersegmentation, three thresholds have been taken into account, for each selected image, to simulate the various situations. Experimental results permit to reveal the advantages and limitations of the studied criteria.

1. INTRODUCTION

Segmentation is one of the first steps in image analysis. It greatly influences the interpretation which will be done afterwards. Many segmentation methods have been proposed in the literature [2]. Each of them lays the emphasis on different properties. This variety makes it difficult to evaluate their efficiency. Actually, many works have been performed to solve the more general problem of the evaluation of image segmentation results [11], [5]. The proposed methods can be classified into two groups. The first one is composed of unsupervised evaluation criteria based on the computation of different statistics that help to quantify the quality of a segmentation result without any *a priori* knowledge [4]. The second one gathers the evaluation methods based upon the computation of a dissimilarity measure between a segmentation result and a ground truth that is either determined by an expert or set during the generation of synthetic images. Even if these methods are inherently dependent on the confidence in the ground truth, they can be widely used for real applications requiring expert evaluation, such as medical applications. This article is devoted to this kind of approach.

We focused on five evaluation criteria which were tested on a selection of hundred images extracted from the ©Corel database. This basis contains images corresponding to different application fields such as medicine, aerial photography, landscape images ... and was completed with experts ground truths [7].

After presenting the tested criteria and some examples of the considered images we produce experimental results of evaluation and compare the efficiency of the various methods.

The authors would like to thank financial support provided by the Conseil Régional du Centre and the European Union (FSE).

2. CONTEXT OF THE STUDY

2.1 Supervised evaluation criteria

Among all the methods proposed in the literature [11], we selected five supervised evaluation criteria :

- Multi-features quality measurement (Quality) [10] : this technique combines the computation of an objective divergence (between the extracted contours and those proposed by the experts) and a subjective evaluation of the different possible errors. Each kind of error is modified by a penalty coefficient that reflects the relative importance attached to this error by the experts. The expression of the quality measurement contains a set of coefficients which are first determined on a benchmark and which can be modified for specific applications and for different experts.
- Pratt's Figure Of Merit (FOM) [9]: this criterion corresponds to an empirical distance between the ground truth composed of contours I_t and those obtained in the chosen segmentation result I_s :

$$FOM(I_t,I_s)=\frac{1}{Max\{card(I_t),card(I_s)\}}\sum_{i=1}^{card(I_s)}\frac{1}{1+d^2(i)} \quad (1)$$

  where $d(i)$ is the distance between the i^{th} pixel of I_s and the nearest pixel of I_t.
- Hausdorff distance (Hausdorff) [1] : this criterion measures the distance between two pixel sets : $I_t = t_1,\ldots,t_m$ and $I_s = s_1,\ldots,s_n$.

$$H(I_t,I_s) = Max(h(I_t,I_s),h(I_s,I_t)) \quad (2)$$

  where

$$h(I_t,I_s) = Max_{t_i\in I_t} min_{s_i\in I_s}||t_i - s_i|| \quad (3)$$

  If $H(I_t,I_s) = d$, this means that all the pixels belonging to I_t are not farther than d from some pixels of I_s. This measure is theoretically very interesting. It indeed gives a good similarity measure between the two images.
- Odet's criteria (ODI_n and UDI_n) [8] : different measurements have recently been proposed to estimate various errors in binary segmentations. Among them, two divergence measures seem to be particularly interesting :

$$ODI_n = \frac{1}{N_o}\sum_{k=1}^{N_o}\left(\frac{d_o(k)}{d_{TH}}\right)^n \quad (4)$$

and

$$UDI_n = \frac{1}{N_u} \sum_{k=1}^{N_u} \left(\frac{d_u(k)}{d_{TH}} \right)^n \quad (5)$$

where

- $d_o(k)$ is the distance between the k^{th} pixel belonging to the segmented contour and the nearest pixel of the reference contour.
- $d_u(k)$ is the distance between the k^{th} non-detected pixel and the nearest one belonging to the segmented contour.
- N_o corresponds to the number of oversegmented pixels.
- N_u corresponds to the number of undersegmented pixels.
- d_{TH} is the maximum distance allowed to search for a contour point.
- n is a scale factor which permits to give a different weight to a pixel depending on its distance from the reference contour.

The ODI_n criterion evaluates the divergence between the oversegmented pixels and the reference contour. The UDI_n criterion estimates the divergence between the undersegmented pixels and the calculated contour.

For nearly all the presented criteria (*Quality*, *Hausdorff*, ODI_n and UDI_n), the lower the criteria, the more efficient the segmentation is. The FOM criterion is an except : the higher the criterion, the more efficient the segmentation is.

2.2 Image database

The database used for our tests includes 100 real images extracted from the ©Corel database for which manual segmentations provided by experts are available. Nine segmentation algorithms have been considered [6] :

- Brightness Gradient (BG).
- Texture Gradient (TG).
- Color gradient (CG).
- Gradient Magnitude (GM).
- Oriented Energy (OE).
- Brightness/Texture Gradients (BGTG).
- Brightness/Color/Texture Gradients (BGCGTG).
- Canny.
- Second Moment Matrix (SMM).

Most of them are based on threshold selection. In order to compare the behavior of the different criteria towards over- and undersegmentation, three thresholds have been taken into account, for each selected image, to simulate the various situations. Figure 1 presents three examples of the selected images, the corresponding ground truths and the three segmentation results obtained with the BGCGTG method. Table 1 presents the corresponding values of the different criteria.

2.3 Discussion

Figure 2 presents the evolutions of the five criteria for 10 images extracted at random from the ©Corel database. The segmentation method we used was once again the BGCGTG method. We can notice that, in nearly all cases, both *Quality* and *Hausdorff* consider the oversegmented situations as the best ones. However, we can observe in table 2 that the results obtained with *Hausdorff* are much more disparate than

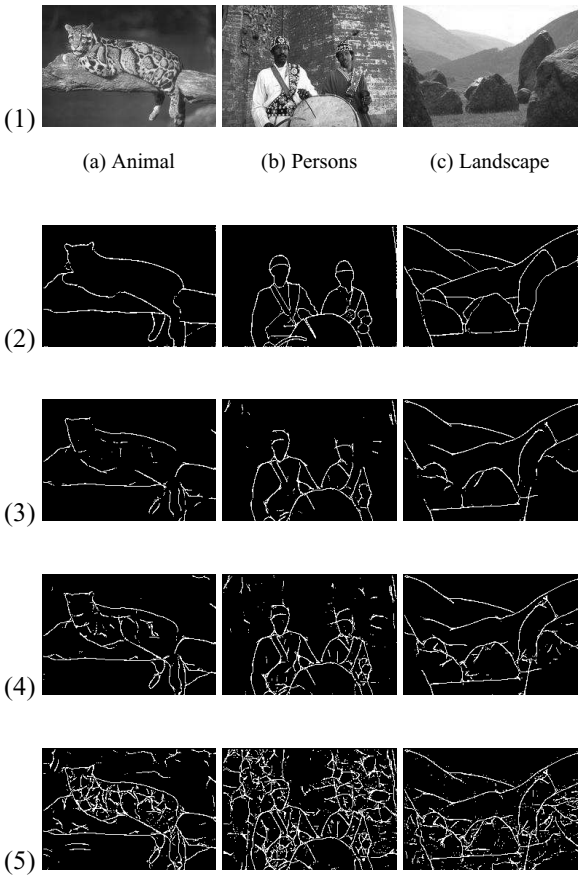


Figure 1: Examples of test images extracted from the ©Corel database. (1) Original images - (2) Corresponding ground truths - BGCGTG segmentation method : (3) undersegmented, (4) normal, (5) oversegmented

		Animal	Persons	Landscape
Quality	underseg.	9152	10512	8223
	normal	8061	8887	7041
	overseg.	**7961**	**8599**	**6658**
FOM	underseg.	0.3737	0.4461	0.5225
	normal	**0.4236**	**0.5381**	**0.5870**
	overseg.	0.2555	0.2641	0.3843
Hausdorff	underseg.	3577	3305	4100
	normal	1865	1685	2993
	overseg.	**1665**	**1125**	**2308**
ODI_n	underseg.	0.7590	**0.6650**	**0.5750**
	normal	0.7520	0.6750	0.6030
	overseg.	**0.7480**	0.6690	0.6390
UDI_n	underseg.	0.8731	0.8534	0.8057
	normal	0.8023	0.7867	0.7565
	overseg.	**0.6527**	**0.4735**	**0.5711**

Table 1: Evaluation criteria for the images presented figure 1.

those obtained with *Quality*. The confidence we can have in Hausdorff's criterion is thus limited. The evolutions obtained with Odet's criteria perfectly match the expected behaviors. ODI_n always ranks first the undersegmented segmentations.

As ODI_n enables us to measure oversegmentation, it logically gives a good mark to undersegmented images. Conversely UDI_n favours the oversegmented situations. Nevertheless, UDI_n seems to be more discriminating. Finally, the FOM criterion stands out. For the 10 presented images, it is the only one which systematically considers that the "normal" segmentation is the best one.

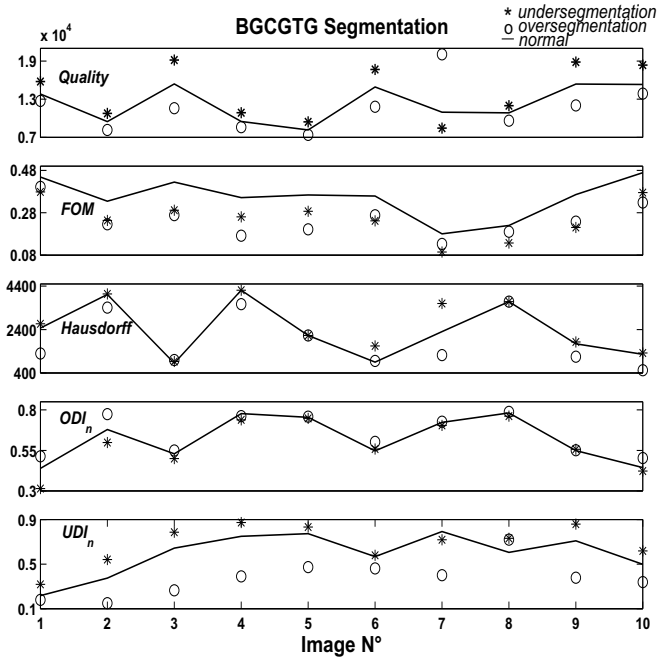


Figure 2: Evolutions of the five criteria face to BGCGTG segmentations.

All these conclusions are confirmed by Table 2 which presents how many times each segmentation (under-, normal and over-) has been considered as the best by each criterion. These statistics highlight the favoured situations for each criterion.

	Undersegmented	Normal	Oversegmented
Quality	7	20	**73**
FOM	0	**99**	1
Hausdorff	7	42	**68**
ODI_n	**49**	37	15
UDI_n	0	10	**90**

Table 2: Global classification, by each criterion, of the different situations (under-, normal or oversegmentation) for 100 images extracted from the ©Corel database and 9 segmentations methods.

For a chosen criterion, one can notice that the sum of the privileged situations sometimes overlaps 100%. This occurs when the different criteria do not distinguish the under-, normal or oversegmentations. We can note that, with *Quality* and *Hausdorff*, oversegmentation is often ranked first (*Quality* even more clearly). As a result of their definitions, ODI_n and UDI_n favour the under- and the oversegmentation respectively. *FOM* again seems to be giving the best results.

All these conclusions are of course highly dependent on the confidence in the ground truth. In fact, as shown in figure 3, different experts can give quite different reference contours for the same image. This therefore relativizes what can be considered as under- or oversegmented and the final conclusion concerning the "best" result.

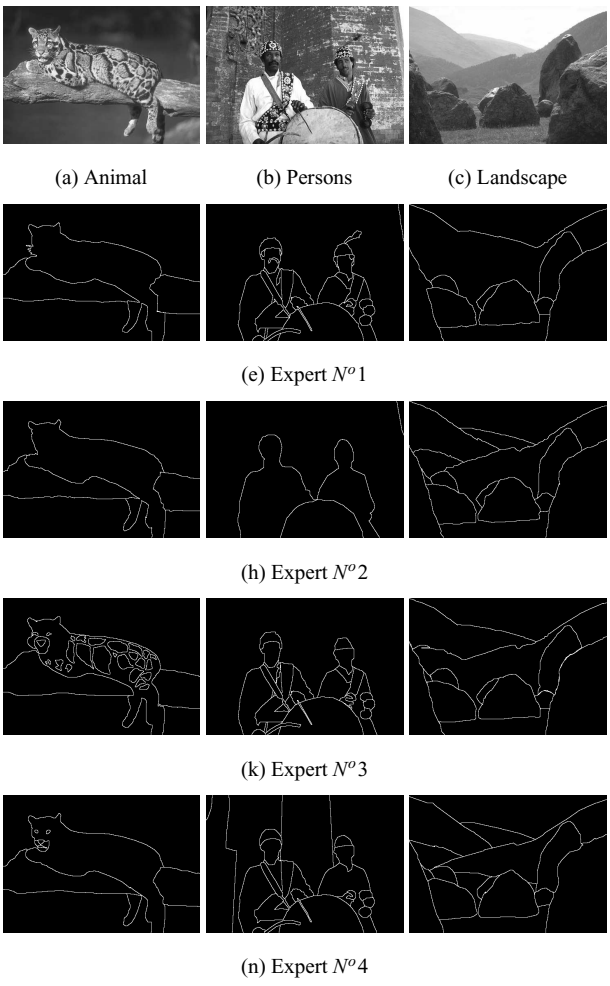


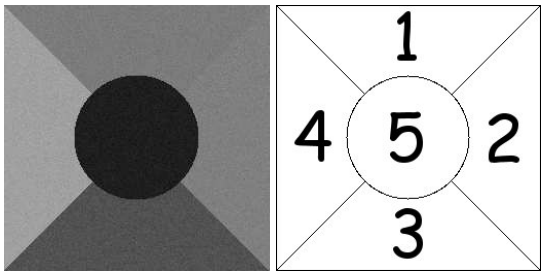
Figure 3: Different reference contours given by four experts.

In order to avoid any subjective interpretation, we finally tested the five criteria on synthetic images for which a completely reliable ground truth is available. In order to illustrate the behaviors of the criteria in this situation, we present in the next paragragraph one example of the tested images.

2.4 Test on a synthetic image

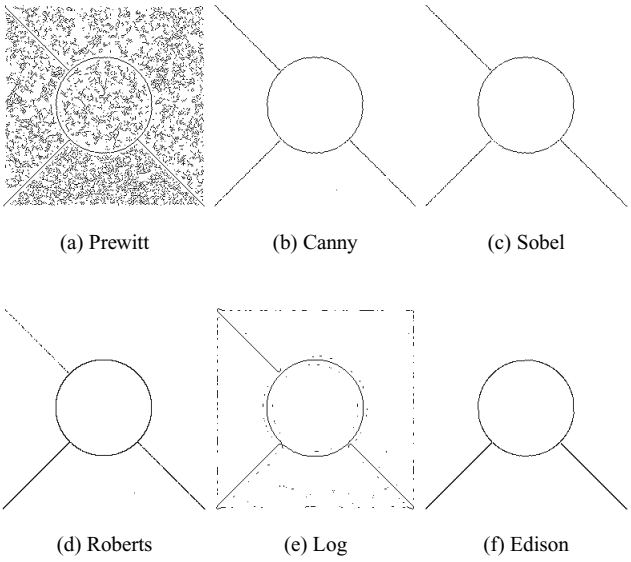
The tested image was composed of five regions that had a high contrast. Only two regions (1 and 2) had quite similar mean values and variances. The image and its ideal segmentation are presented figure 4. Figure 5 presents six segmentation results and table 3 gathers, for each segmentation, the values of the five considered criteria.

UDI_n and *Hausdorff* once again favour the oversegmen-



(a) (b)

Figure 4: Synthetic image (a) and corresponding ground truth (b)



(a) Prewitt (b) Canny (c) Sobel

(d) Roberts (e) Log (f) Edison

Figure 5: Segmentation results of figure 4(a)

	Quality	FOM	Hausdorff	ODI_n	UDI_n
Prewitt	5477	0.0548	4513	0.945	**0.0166**
Canny	2508	0.6759	14450	0.004	0.2276
Sobel	**2446**	0.6875	14450	**0.001**	0.2346
Roberts	3177	0.6696	14450	0.004	0.2251
Log	4520	0.3483	**4346**	0.450	0.1389
Edison	3629	**0.7643**	14450	**0.001**	0.5212

Table 3: Evalution criteria for the different segmentations.

tations (*Log* and *Prewitt*), while conversely ODI_n gives a good mark to undersegmented situations (*Edison* [3]). The *FOM* criterion considers the *Edison* segmentation as the best one. The next ones are *Sobel* and *Canny*. This is probably due to the very precise contours detected by *Edison* even if a frontier is missing. On the other hand, *Quality* is the only criterion that seems to have a behavior which is completely different from its behavior in the previous analysis. It indeed doesn't favour oversegmentations anymore. This can be explained by the importance of a good appropriateness between the set coefficients that intervene in the definition and the context of the study. This criterion is no longer suitable if we change the application without adapting the coefficients.

3. CONCLUSION AND PERSPECTIVES

In this article we have presented a comparison of five supervised evaluation criteria of segmentation results. One criterion stands out from this study : Pratt's Figure of Merit (*FOM*). With real and synthetic images corresponding to different application fields, this criterion revealed itself as the most effective.

Moreover two criteria, ODI_n and UDI_n, have proved their efficiency to measure under- and oversegmentation respectively. It could be interesting to combine these two functions in order to obtain a reliable evaluation criterion.

REFERENCES

[1] M. Beauchemin, KP.B. Thomson, G. Edwards, "On the Hausdorff distance used for the evaluation of segmentation results," *CJRS*, vol. 24(1), pp. 3–8, 1998.

[2] R. Chellappa, B.S. Manjunath, "Texture Classification and Segmentation," *FIU01*, Chapter 8.

[3] D. Comanicu, P. Meer, "Mean shift : A robust approach toward feature space analysis," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 24, pp. 603–619, 2002.

[4] H. Laurent, S. Chabrier, C. Rosenberger, B. Emile, P. Marché, "Etude comparative de critères d'évaluation de la segmentation," in *Proc. GRETSI 2003*, Paris, France, Sept. 2003, vol. 3 pp. 150–153.

[5] V. Letournel, "Contribution à l'évaluation d'algorithmes de traitements d'images," *PhD thesis*. ENST Paris, Dec. 2002.

[6] D. Martin, C. Fowlkes, J. Malik, "Learning to detect natural image boundaries using local brightness, color and texture cues," in *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 2004.

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

[8] C. Odet, B. Belaroussi, H. Benoit-Cattin, "Scalable discrepancy measures for segmentation evaluation," in *Proc. ICIP 2002*, vol. 1, pp. 785–788.

[9] W. Pratt, O.D. Faugeras, A. Gagalowicz, "Visual discrimination of stochastic texture fields," *IEEE Trans. on Systems, Man and Cynernetics*, vol. 11, pp. 796–804, 1978.

[10] R. Roman-Roldan, J.F. Gomez-Lopera, C. Atae-Allah, J. Martinez-Aroza, P.L. Luque-Escamilla, "A measurement of quality for evaluating methods of segmentation and edge detection," *Pattern Recognition*, vol. 34, pp. 969–980, 2001.

[11] Y.J. Zhang, "A survey on evaluation methods for image segmentation," *Computer Vision and Pattern Recognition*, vol. 29, pp. 1335–1346, 1996.