EVALUATING THE SEGMENTATION RESULT OF A GRAY-LEVEL IMAGE

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

Laboratoire Vision et Robotique, UPRES EA 2078
ENSI de Bourges / Université d'Orléans
10 boulevard Lahitolle, 18020 Bourges Cedex, France
phone: +33 2248484000, fax: +33 2248484000, email: sebastien.chabrier@ensi-bourges.fr

ABSTRACT

We propose in this communication an unsupervised criterion that enables to quantify the quality of a segmentation result of a gray-level image. The originality of this method lies in the possibility to evaluate the segmentation results for all kinds of images including textured ones. This method is based upon a new criterion that takes into account the intra-region and inter-regions disparities by considering the type of region (textured or uniform). These two disparity measures are computed from adaptive attributes calculated from the segmentation result. Experimental results show the efficiency of this technique for synthetic and natural images with textured regions.

1. INTRODUCTION

Segmentation is an essential stage in image processing since it conditions the quality of interpretation. Several segmentation methods have been developed and it is difficult to evaluate their efficiency. Actually, many works have been performed to solve the more general problem of the evaluation of a segmentation result and some literature is available. Briefly stated, there are two main approaches.

On the one hand, there are supervised evaluation methods based on the computation of a dissimilarity measure between a segmentation result and the ground truth. These methods are widely use for medical applications. Baddeley's distance [1], Vinet's measure [7] or classification rate are examples of such methods. On the other hand, there are unsupervised evaluation criteria that enable to quantify the quality of a segmentation result by considering different statistics. In [11], we can find a relatively complete list of such methods. Most of these methods are not adapted for texture segmentation results [2].

In order to try to solve this problem, we propose in this communication a new quantitative evaluation criterion. The originality of this method lies in the possibility to evaluate the segmentation results of textured images without any *a priori* knowledge. This criterion combines the intra-region and inter-regions disparities. The whole processing is adapted by considering the determined type of each region (textured or uniform) in the segmentation result.

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

2. PROPOSED CRITERION

There are few criteria which enable to quantify the quality of a segmentation result without any *a priori* knowledge such as the ground truth [11], [8]. Most of these methods have the disadvantage of not being powerful enough for textured images.

The definition of the evaluation criterion we propose is based on the definition of segmentation given by Haralick [5]. A "good" segmentation must satisfy the following two properties:

- A region of the segmented image has to contain a single primitive (a texture or a constant gray scale) to guarantee that there is no under-segmentation. Thus, a region is characterized by the stability of statistics.
- Two adjacent regions have to contain two different primitives to guarantee that there is no over-segmentation. This corresponds to a disparity of statistics between these two regions [11].

2.1 Definitions

Let there be a set of n image segmentation methods called $M = \{M^1, .., M^n\}$. The segmentation result of the image I with the M^j method is designated by I^j.

The evaluation criterion F we propose gives a normalized quantitative value of a segmentation result :

$$F(I^j) = \frac{\overline{D}(I^j) - \underline{D}(I^j)}{2} \tag{1}$$

The global intra-region disparity $\underline{D}(I^j)$ quantifies the homogeneity of each region in the segmented image I^j. Similarly, the global inter-regions disparity $\overline{D}(I^j)$ measures the disparity between the regions. Note that these two disparities are normalized and are computed by taking into account the type of region (uniform or textured). A segmentation result I^k is regarded as better than another one I^l, if we have $F(I^k) > F(I^l)$.

The number of regions composing the segmentation result obtained from the M^j method may vary depending on the used method. We define m^j as the number of R_i^j regions obtained from the M^j method, r_i^j as the number of pixels in the region R_i^j and NT as the total number of pixels in the image I.

The global intra-region disparity $\underline{D}(I^j)$ is computed from the intra-region disparity of each region $\underline{D}(R_i^j)$:

$$\underline{D}(I^j) = \frac{1}{m^j} \sum_{i=1}^{m^j} \frac{r_i^j}{NT} \underline{D}(R_i^j) \tag{2}$$

The global intra-region disparity is ponderated according to the number of r_i^j pixels of each region R_i^j. The larger a region is, the more influence it has in the global intra-region disparity. The computation of the global inter-regions disparity $\overline{D}(I^j)$ is similar.

We must first determine the textured and the uniform regions in order to compute this criterion for each region.

2.2 Texture and Uniform region detection

If we refer to literature, the textural character of an image or a region is usually determined by considering the standard deviation of the gray scale [10]. In this case, a region of an image is set as textured if its standard deviation is above a defined threshold. Nevertheless, the standard deviation does not take into account the distribution of gray scale but only their variations in relation to the average gray scale of a region. In fact, texture depends on the scale of observation so that two images with the same standard deviation can be textured for one and uniform for the other (composed of large uniform regions).

In order to determine the textural character of a region, we defined a parameter called "uniformity" derived from the cooccurrence matrix [9]. This parameter noted U corresponds to the average of the diagonal entries in the cooccurrence matrix for one unit displacement distance and four orientations.

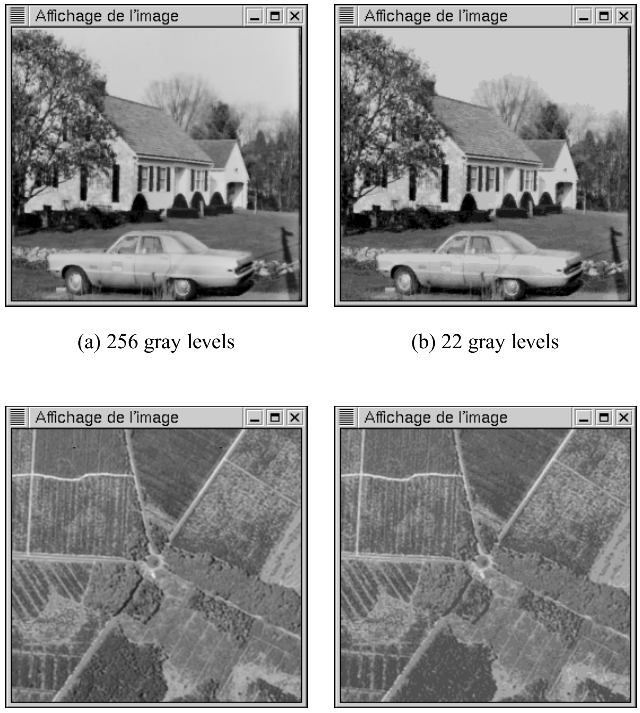
The cooccurrence matrix is a time-consuming process for an image with several gray levels. In order to decrease the number of gray levels while preserving the meaningful information, the image is thresholded. The thresholding method used [6] enables the selection of the most significant gray levels after a local analysis of the image. This approach consists in sorting the pixels of the image by comparing their gray levels with a threshold computed from transformed local histograms. The transformation used allows the distinction of the principal modes of each histogram by using four criteria that aim at reproducing the eyes' sensitivity to contrast [3]. To show the efficiency of this method, we present, in figure 1, two examples of thresholded images together with the corresponding original images.

Parameter U, calculated in region R, gives an indication on the global character of the uniformity of a region. In order to identify the global nature of a region, we defined the following decision criterion:

$$\begin{cases} if\ U(R) < \tilde{U}(R) & \text{then the region is mainly textured} \\ otherwise & \text{the region is mainly uniform} \end{cases} \tag{3}$$

where

$$\tilde{U}(R) = 1 - \left(\frac{NGR-1}{NGR}\right)^8 \tag{4}$$



(a) 256 gray levels (b) 22 gray levels

(c) 256 gray levels (d) 13 gray levels

Figure 1: Thresholding results of two images

where NGR is the number of gray levels in the region. The value of parameter $\tilde{U}(R)$ represents the probability of transition of gray scale of a pixel within an 8 connected neighborhood under the hypothesis of independence of gray scale.

2.3 Intra-region disparity

The intra-region disparity $\underline{D}(R_i^j)$ of the region R_i^j is computed considering its type. A region containing two different primitives must have a high intra-region disparity compared to the same region composed of a single primitive.

- Uniform case

The intra-region disparity, in the *uniform case*, is equal to the normalized standard deviation of the region (which is here a sufficient piece of information).

- Textured case

In the *textured case*, each region is characterized by a set of texture attributes vector (computed from a sliding window). The dispersion of this set of vectors allows us to calculate the intra-region disparity in the textured case.

2.4 Inter-regions disparity

The inter-regions disparity of two neighboring regions is also computed by taking into account their types.

- Regions of same type

a) Uniform regions

This parameter is computed as the average of the disparity of a region with its neighbors. The disparity of *two uniform regions* R_i and R_j is calculated as:

$$\overline{D}(R_i,R_j) = \frac{|E[R_i] - E[R_j]|}{NGR} \tag{5}$$

where $E[R_i]$ is the average gray scale in the region R_i.

b) Textured regions
The disparity of *two textured regions* R_i and R_j is defined as:

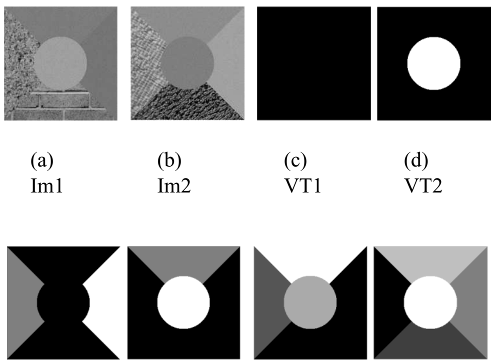
$$\overline{D}(R_i,R_j) = \frac{d(B_i,B_j)}{||B_i|| + ||B_j||} \tag{6}$$

where $d(.,.)$ is the Euclidean distance, $||B_i||$ is the quadratic norm of the average value of adaptive attributes characterizing region R_i.

- Regions of different types
The disparity of *regions of different types* is set as the maximal value 1.

3. EXPERIMENTAL RESULTS

First of all, we evaluated 6 synthetic segmentation results of 100 images so as to validate the proposed criterion. We show in table 1 the evaluation results for the different synthetic segmentation results of the images presented in figure 2(a) and (b). For each image, the segmentation result $VT5$ is the best one as shown by the value of the proposed criterion. Note that the value of the criterion usually increases with the number of correct classes except once.



(a) Im1 (b) Im2 (c) VT1 (d) VT2

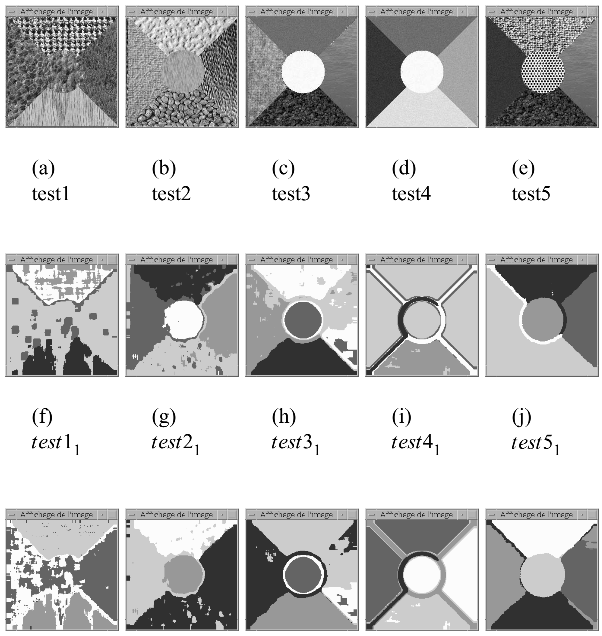
(e) VT3 (f) VT3b (g) VT4 (h) VT5

Figure 2: Synthetic segmentation results of two original images

Image	$VT1$	$VT2$	$VT3$	$VT3b$	$VT4$	$VT5$
image1	0	0.51	0.59	0.61	0.66	0.84
image2	0	0.25	0.39	0.60	0.52	0.64

Table 1: Evaluation criterion of synthetic segmentation results of Figure 2

In order to measure the efficiency of the proposed evaluation criterion, we give the segmentation results of several gray scale images with a set ground truth (see Figure 3). These images are composed of textured and uniform regions. The segmentation results $testi_1, i = 1..5$ were obtained by using the K-means algorithm with gray scale moments of order 1 to 4. The segmentation results $testi_2, i = 1..5$ were obtained by using the K-means algorithm with 15 statistical features from the cooccurrence matrix. Table 2 shows the value of the proposed evaluation criterion and the correct classification rate for each segmentation result. These results show that the variation of the evaluation criterion is coherent. Indeed, if a segmentation result is better as another ones if we consider the correct classification rate, it is also the case then for our evaluation criterion.



(a) test1 (b) test2 (c) test3 (d) test4 (e) test5

(f) $test1_1$ (g) $test2_1$ (h) $test3_1$ (i) $test4_1$ (j) $test5_1$

(k) $test1_2$ (l) $test2_2$ (m) $test3_2$ (n) $test4_2$ (o) $test5_2$

Figure 3: 10 segmentation results of 5 test images (K-means with texture attributes)

Result	$\underline{D}(I)$	$\overline{D}(I)$	F	correct classification
$test1_1$	0.26	0.92	0.33	41.6%
$test1_2$	0.04	0.80	0.38	58.2%
$test2_1$	0	0.61	0.30	86.2%
$test2_2$	0.28	0.63	0.18	73%
$test3_1$	0.31	0.81	0.25	78.8%
$test3_2$	0.68	0.8	0.01	54.6%
$test4_1$	0.52	0.67	0.07	20.8%
$test4_2$	0.37	0.59	0.11	44%
$test5_1$	0.58	0.95	0.19	75%
$test5_2$	0.16	0.73	0.29	93.8%

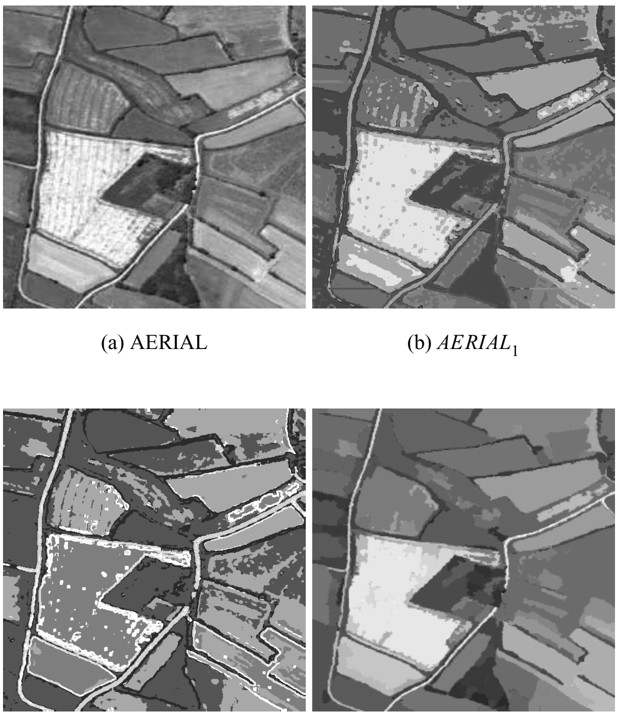
Table 2: Comparison between the evaluation criterion and the correct classification rate of segmentation results as shown in Figure 3

We compared the proposed criterion with different evaluation criteria in the literature [11] : Borsotti, Zeboudj, Inter-regions, Intra-regions. We used a database of 100 images, with 4 segmentation results for each one. We compared the values of these criteria with Vinet's criterion (classification rate). Table 3 gives the correlation factor of each criterion. One can see that the maximal absolute value of the correlation factor of the different criteria with Vinet's measure (used here as an objective measure) is obtained with the proposed criterion.

Borsotti	1	-0.09	-0.03	0.77	-0.07	**-0.07**
Zeboudj		1	0.81	0.02	-0.06	**-0.21**
Inter-region			1	0.06	-0.03	**-0.01**
Intra-region				1	-0.17	**-0.13**
Proposed					1	**0.25**
Vinet						**1**

Table 3: Correlation factors of the different criteria

We applied the proposed criterion in order to evaluate three segmentation results of an image without any ground truth (see Figure 4.). The first segmentation result is obtained with the method proposed in [9], the second one with the K-means algorithm by using the mean and variance as attributes and the last one with the EDISON algorithm [4]. Table 4 shows that the value of the criterion is also coherent if we consider the visual perception of a segmentation result. The evaluation of each segmentation result (of size 256×256 pixels) took 250 seconds on a ©Pentium 4 with 3Ghz.



(a) AERIAL (b) $AERIAL_1$

(c) $AERIAL_2$ (d) $AERIAL_3$

Figure 4: Three segmentation results of image "AERIAL"

Result	F
$AERIAL_1$	0.598
$AERIAL_2$	0.506
$AERIAL_3$	**0.603**

Table 4: Evaluation criterion of segmentation results of image "AERIAL"

4. CONCLUSION

The proposed method enables to quantify the quality of a segmentation result without any *a priori* knowledge for all kinds of gray-level images including textured ones. The defined criterion combines the intra-region and inter-regions disparities of regions in the segmentation result by taking into account their types. This technique allows us to evaluate the efficiency of a segmentation method for different types of images [1].

Prospects for this study concern the generalisation of the defined criterion to multi-component images.

REFERENCES

[1] A. J. Baddeley. An error metric for binary images. *Robust Computer Vision*, pages 59–78, 1992.

[2] K.A. Bartels and J.L. Fisher. Multifrequency eddy current image processing techniques for nondestructive evaluation. In *ICIP*, pages 486–489, 1995.

[3] K. Chehdi C. Kermad. Multi-bands image segmentation : A scalar approach. *IEEE International Conference on Image Processing*, (6):1012–1025, 2000.

[4] D. Comanicu and P. Meer. Mean shift: A robust approach toward feature space analysis. *IEEE Trans. Pattern Anal. Machine Intell.*, 24:603–619, 2002.

[5] R.M. Haralick. Statistical and structural approaches to texture. In *proceedings of the IEEE*, volume 69, pages 786–804, May 1979.

[6] C. Kermad, B. Vozel, and K. Chehdi. Hyperspectral image analysis and dimensionality : a scalar scheme through multi-thresholding technique. *Proc of the Eos/Spie Symposium on Remote sensing*, 31(4170), 2000.

[7] Vinet L. *Segmentation et mise en correspondance de regions de paires d'images stereoscopiques*. PhD thesis, Universite de Paris IX Dauphine, July 1991.

[8] R. Roman-Roldan, J.F. Gomez-Lopera, C. Atae-Allah, J. Martinez-Aroza, and P.L. Luque-Escamilla. A measure of quality for evaluating methods of segmentation and edge detection. *Pattern Recognition*, (34):969–980, 2001.

[9] C. Rosenberger and K. Chehdi. Towards a complete adaptive analysis of an image. *Electronic Imaging*, 2(2):292–298, 2003.

[10] C. S. Won and Y. Choe. Image block classification using stochastic image segmentation. *Electronics Letters*, 32(16):1462–1463, August 1996.

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

[1] The proposed evaluation criterion is available at http://www.ensi-bourges.fr/perso/chabrier/telechargements.htm