Measuring the Quality Evaluation for Image Segmentation

Rodrigo Janasievicz Gomes Pinheiro and Jacques Facon

PUCPR-Pontifícia Universidade Católica do Paraná,
Rua Imaculada Conceição, 1155, Prado Velho,
80215-901 Curitiba-PR, Brazil
{pinheiro, facon}@ppgia.pucpr.br

Abstract. This paper proposes a measure of quality for evaluating the performance of region-based segmentation methods. Degradation mechanisms are used to compare segmentation evaluation methods onto deteriorated ground-truth segmentation images. Experiments showed the significance of using degradation mechanisms to compare segmentation evaluation methods. Encouraging results were obtained for a selection of degradation effects.

1 Introduction

Image Segmentation is a field that deals with the analysis of the spatial content of an image. It is used to separate semantic sets (regions, textures edges) and is an important step for image understanding. The region-based segmentation consists in estimating which class each pixel of the image belongs to. Due to the fact that none of the segmentation approaches are applicable to all images, several region-based segmentation approaches have been proposed. None of the algorithms are equally suitable for a particular application. It is the reason why establish certain criteria, other than human subjective ones, to evaluate the performance evaluation of segmentation algorithms is needed. Performance evaluation is a critical step for increasing the understanding rates in image processing. This work will focus on discrepancy evaluation methods of region-based segmentation, that consist in comparing the results obtained by applying a segmentation algorithm with a reference (ground-truth) and measuring the differences (or discrepancy). Zhang [1] has proposed a discrepancy evaluation method based on mis-classified pixels. Suppose an image segmented into N pixel classes, a confusion matrix C of dimension $N \times N$ can be constructed, where each entry C_{ij} represents the pixel number of class j classified as class i by the segmentation algorithms. A first error type named ”multi-class Type I error” was defined as:

$$M_I^{(k)} = 100 \times \left[\left(\sum_{i=1}^{N} C_{ik} \right) - C_{kk} \right] / \left[\sum_{i=1}^{N} C_{ik} \right] \tag{1}$$

where the numerator represents the pixel number of class k not classified as k and the denominator is the total pixel number of class k. A second error type named ”multi-class Type II error” was defined as:

M. Lazo and A. Sanfeliu (Eds.): CIARP 2005, LNCS 3773, pp. 120–127, 2005.

$$M_{II}^{(k)} = 100 \times \left[\left(\sum_{i=1}^{N} C_{ki} \right) - C_{kk} \right] / \left[\left(\sum_{i=1}^{N} \sum_{j=1}^{N} C_{ij} \right) \right] \tag{2}$$

where the numerator represents the pixel number of other classes called class k. The denominator is the total pixel number of other classes.

Yasnoff et al [2] have shown that measuring the discrepancy only on the number of mis-classified pixels does not consider the pixel position error. Possible solution is to use the distance between the mis-segmented pixel and the nearest pixel that actually belongs to the mis-segmented class. Let S be the number of mis-segmented pixels for the whole image and $d(i)$ be a metric to measure the distance between the i^{th} mis-segmented pixel and the nearest pixel that actually is of the mis-classified class. Yasnoff et al [2] have defined a discrepancy measure D based on this distance:

$$D = \sum_{i=1}^{S} d^2(i) \tag{3}$$

To exempt the influence of image size, the discrepancy measure D is normalized ND:

$$ND = 100 \times \sqrt{D} \Big/ T \tag{4}$$

where T is the total pixel number in the image.

This work will focus on proposing a new discrepancy evaluation and a strategy for measuring its performance. This paper is organized as follows: A new discrepancy evaluation method taking into account the different "scenarios" occurred in a segmentation process is detailed in Section 2. Section 3 presents some experimental results and discussions. Section 4 shows the quality evaluation of two specific address block segmentation methods. Finally, some conclusions are drawn in Section 5.

2 New Discrepancy Evaluation Method

A discrepancy evaluation method taking into account the different "scenarios" occurred in a segmentation process is proposed. Let A be a segmentation algorithm to be evaluated. Let G_i (where i = 1 to G) be the ground-truth regions of a image and S_j (where j = 1 to S) be the segmented regions obtained from algorithm A. Let n_{Gi} be the pixel number of the ground-truth region G_i, and n_{Sj} be the pixel number of the segmented region S_j. Let also $w_{ij} = n_{Gi} \cap n_{Sj}$ be the number of well-classified pixels between regions G_i e S_j. A discrepancy measure D_i is defined for each ground-truth region. G_i. To characterize the discrepancy between G_i and S_j, four classifications of region segmentation are considered:

- Correct segmentation: The ground-truth region G_i has been segmented in an unique region S_j: the discrepancy measure is $D_i = w_{ij}$. In case of total overlap, $D_i = w_{ij} = n_{Gi} = n_{Sj}$.
- Over-segmentation: The segmentation process has fragmented the ground-truth region G_i in a set of s regions S_j: the discrepancy measure is $D_i = w_{ij}/s$;

- Under-segmentation: The segmentation process has merged a set of g ground-truth region G_i in an unique region S_j: the discrepancy measure is $D_i = w_{ij}/g$;
- mis-segmentation: The ground-truth region G_i has not been segmented. The discrepancy measure in this case represents a penalty: $D_i = -n_{Gi}$;

A general metric $\Upsilon(A)$, taking into account these four ”scenarios”, can qualify the segmentation method A, as follows:

$$\Upsilon(A) = \frac{\sum_{i=1}^{G} D_i}{\sum_{i=1}^{G} n_{Gi}} \tag{5}$$

This metric $\Upsilon(A)$ presents some properties:

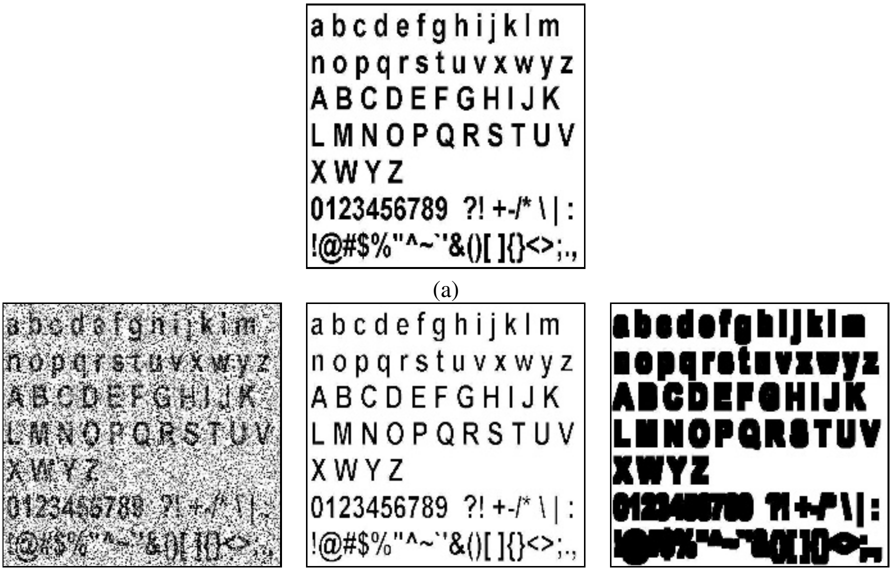
- $\Upsilon(A) = -1$ when segmentation totally failed (the A algorithm has ignored all ground-truth regions);
- $\Upsilon(A) = 0$ when the number of correct, over or under-segmented pixels matches the number of ”forgotten” pixels;
- $\Upsilon(A) = 1$ when segmentation has completely succeeded;
- Metric $\Upsilon(A)$ verifies $-1 \leq \Upsilon(A) \leq 1$.

3 Comparison Strategy and Results

In order to compare different segmentation methods, two strategies can be used: the first one consists in applying the evaluation methods to segmented images obtained from different segmentation approaches. The second one consists in simulating results of segmentation processes. The latter has been adopted and a set of test images synthetically deteriorated was used. A binary image (640 × 480 pixels) that represents the ground-truth segmentation has suffered deteriorations. By this way, the aim is evaluating the resistance of segmentation methods to noise, shrinking and stretching. The degradation processes are a combination of salt noise, pepper noise and salt-pepper noise ($\{1, 5, 10, 15, 20, 25, 30, 35, 40, 45, 50\%\}$), i ($\in [1, 5]$) erosions and dilations (both with cross and square structuring elements EE. Fig 1 depicts some of these test images used during the evaluation process.

Five discrepancy criteria have been applied to a database of 90 deteriorated images (Fig 1) where image 1-(a) represents the ground-truth image: The Zhang [1] multi-class Types I and II error criteria (equations 1 and 2) , the Yasnoff et al [2] discrepancy measure ND (equation 4), the new proposed evaluation metric (equation 5), respectively named $DBMCP-Type\ I$, $DBMCP-Type\ II$, $DBSMSP$ and $\Upsilon(.)$. By modifying the ND measure, a fifth discrepancy measure , named $DBSMSP-II$, has been used, where $d(i)$ measures the distance between the i^{th} mis-segmented pixel and the gravity center of the nearest ground-truth class.

For the aim of comparison, results are depicted in Figures 2, 3, 4. $DBMCP - Type\ I$, $DBMCP - Type\ II$, $DBSMSP$, $DBSMSP - II$ measures have been inverted and $\Upsilon(.)$ metric normalized between 0 and 1.



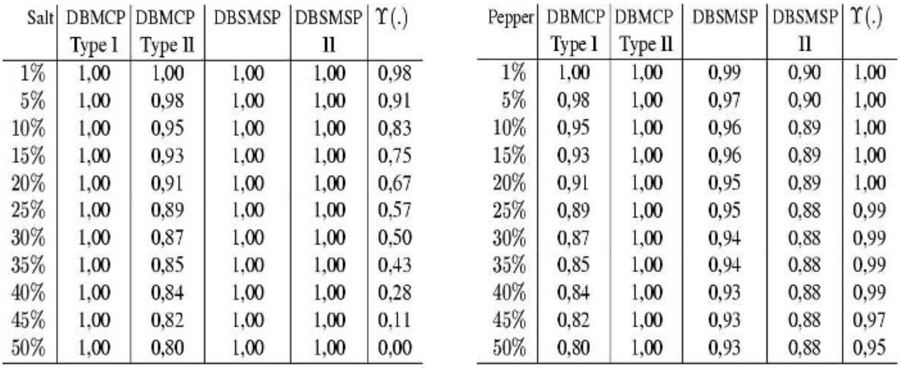
(a)

(b) (c) (d)

Fig. 1. Test images: (a) Ground-truth image, (b) Salt-Pepper (50%), (c) Dilation (cross EE 1 iteration), (d) Erosion (square EE 5 iterations)

It may be observed that:

- The $DBMCP - Type\ I$, $DBSMSP$ and $DBSMSP - II$ measures are totally insensitive with respect to ”holed” segmentation simulated from salt noise (Figure 2-(a). The $DBMCP - Type\ II$ measure is few sensitive. In the opposite, the new $\Upsilon(.)$ criterion is very sensitive to the ”salt” effect;
- No measure is really sensitive with respect to noisy segmentation simulated from pepper noise (Figure 2-(b);
- With respect to black set expansion simulated from dilation, $DBMCP - Type\ II$ and $\Upsilon(.)$ criteria is very sensitive (Figure 3-(b));
- With respect to black set shrinking simulated from erosion, no measure (Figure 3-(a)) shows high sensibility. The $DBMCP - Type\ I$ measure is the more sensitive;
- With respect to black set expansion and salt-pepper noise, all measures (Figure 4-(a)) show low sensibility. The proposed metric $\Upsilon(.)$ is less sensitive than other ones when erosion is combined with few salt-pepper. On the other hand, with increasing salt-pepper percent, the metric $\Upsilon(.)$ decreases faster than other ones;
- With respect to bad segmentation simulated from dilation and salt-pepper noise, $DBSMSP - II$, $DBMCP - Type\ I$ and $DBSMSP$ measures are not very sensitive (Figure 4-(b)). In case of serious degradation (one EE_{cross} dilation and 50% of salt-pepper noise), these criteria do not decrease below 77%. $DBMCP$ $-Type\ II$ criterion is a little bit more sensitive and does not decrease below 80%. The new $\Upsilon(.)$ criterion is much more robust and reliable in evaluating this kind of bad segmentation.



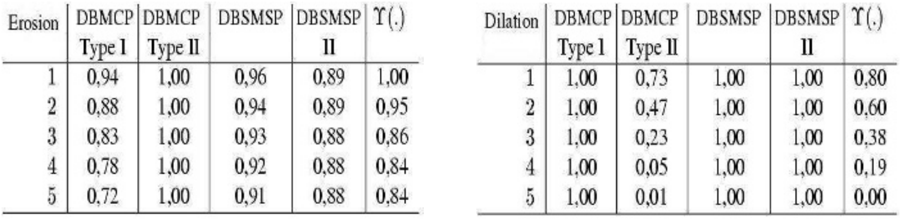
Salt	DBMCP Type I	DBMCP Type II	DBSMSP	DBSMSP II	$\Upsilon(.)$
1%	1,00	1,00	1,00	1,00	0,98
5%	1,00	0,98	1,00	1,00	0,91
10%	1,00	0,95	1,00	1,00	0,83
15%	1,00	0,93	1,00	1,00	0,75
20%	1,00	0,91	1,00	1,00	0,67
25%	1,00	0,89	1,00	1,00	0,57
30%	1,00	0,87	1,00	1,00	0,50
35%	1,00	0,85	1,00	1,00	0,43
40%	1,00	0,84	1,00	1,00	0,28
45%	1,00	0,82	1,00	1,00	0,11
50%	1,00	0,80	1,00	1,00	0,00

(a)

Pepper	DBMCP Type I	DBMCP Type II	DBSMSP	DBSMSP II	$\Upsilon(.)$
1%	1,00	1,00	0,99	0,90	1,00
5%	0,98	1,00	0,97	0,90	1,00
10%	0,95	1,00	0,96	0,89	1,00
15%	0,93	1,00	0,96	0,89	1,00
20%	0,91	1,00	0,95	0,89	1,00
25%	0,89	1,00	0,95	0,88	0,99
30%	0,87	1,00	0,94	0,88	0,99
35%	0,85	1,00	0,94	0,88	0,99
40%	0,84	1,00	0,93	0,88	0,99
45%	0,82	1,00	0,93	0,88	0,97
50%	0,80	1,00	0,93	0,88	0,95

(b)

Fig. 2. Normalised measure values for:(a) Salt noise, (b) Pepper noise



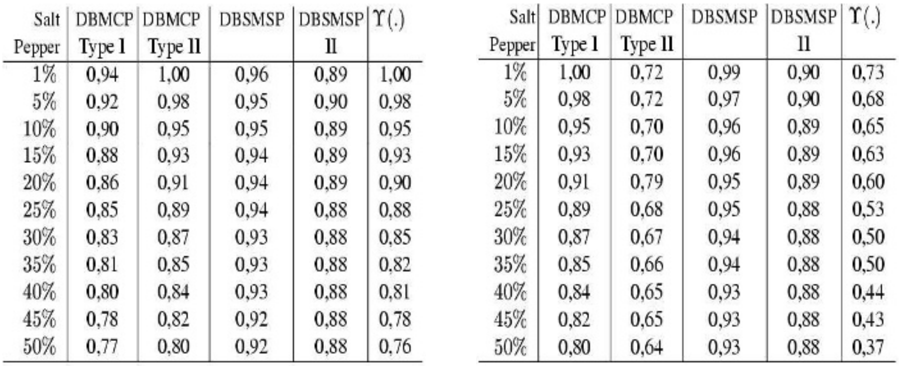
Erosion	DBMCP Type I	DBMCP Type II	DBSMSP	DBSMSP II	$\Upsilon(.)$
1	0,94	1,00	0,96	0,89	1,00
2	0,88	1,00	0,94	0,89	0,95
3	0,83	1,00	0,93	0,88	0,86
4	0,78	1,00	0,92	0,88	0,84
5	0,72	1,00	0,91	0,88	0,84

(a)

Dilation	DBMCP Type I	DBMCP Type II	DBSMSP	DBSMSP II	$\Upsilon(.)$
1	1,00	0,73	1,00	1,00	0,80
2	1,00	0,47	1,00	1,00	0,60
3	1,00	0,23	1,00	1,00	0,38
4	1,00	0,05	1,00	1,00	0,19
5	1,00	0,01	1,00	1,00	0,00

(b)

Fig. 3. Normalized measure values for:(a) Erosion, (b) Dilation



Salt Pepper	DBMCP Type I	DBMCP Type II	DBSMSP	DBSMSP II	$\Upsilon(.)$
1%	0,94	1,00	0,96	0,89	1,00
5%	0,92	0,98	0,95	0,90	0,98
10%	0,90	0,95	0,95	0,89	0,95
15%	0,88	0,93	0,94	0,89	0,93
20%	0,86	0,91	0,94	0,89	0,90
25%	0,85	0,89	0,94	0,88	0,88
30%	0,83	0,87	0,93	0,88	0,85
35%	0,81	0,85	0,93	0,88	0,82
40%	0,80	0,84	0,93	0,88	0,81
45%	0,78	0,82	0,92	0,88	0,78
50%	0,77	0,80	0,92	0,88	0,76

Salt Pepper	DBMCP Type I	DBMCP Type II	DBSMSP	DBSMSP II	$\Upsilon(.)$
1%	1,00	0,72	0,99	0,90	0,73
5%	0,98	0,72	0,97	0,90	0,68
10%	0,95	0,70	0,96	0,89	0,65
15%	0,93	0,70	0,96	0,89	0,63
20%	0,91	0,79	0,95	0,89	0,60
25%	0,89	0,68	0,95	0,88	0,53
30%	0,87	0,67	0,94	0,88	0,50
35%	0,85	0,66	0,94	0,88	0,50
40%	0,84	0,65	0,93	0,88	0,44
45%	0,82	0,65	0,93	0,88	0,43
50%	0,80	0,64	0,93	0,88	0,37

Fig. 4. Normalized measure values with salt-pepper noise for: (a) Erosion, (b) Dilation

4 Real Application and Discussion

In order to test the accuracy of quality evaluation in real segmentation, the five discrepancy criteria were applied on two published approaches for postal envelopes; the first one based on feature selection in wavelet space [3] and the second one based on fractal dimension [4]. In both approaches, the same database composed of 200 complex postal envelope images, with no fixed position for the handwritten address blocks, postmarks and stamps was used. The authors have also employed a ground-truth strategy where the accuracy was computed by only taking in account the identical pixels at the same location.

According to [3], the wavelet-based segmentation rates are 97.36% for address block, 26.96% for stamps and 75.88% for postmarks. According to [4], the fractal-based approach rates are 97.24% for address block, 66.34% for stamps and 91.89% for postmarks.

By applying the five discrepancy criteria to [3] 's and [4] 's segmentation results, without separating the address block, stamp and postmark classes, we obtained the quality evaluation rates grouped in Table 1. This Table depicts that $DBMCP - Type\ I$ and $DBSMSP$ and $DBSMSP - II$ measures have the same sensibility than [3] 's and [4] 's address block evaluation. This fact means that these 3 measures were not able to accurately evaluate the results of real segmentation. The 3 measures have not evaluated that the stamp and postmark segmentation was worse than the address block one.

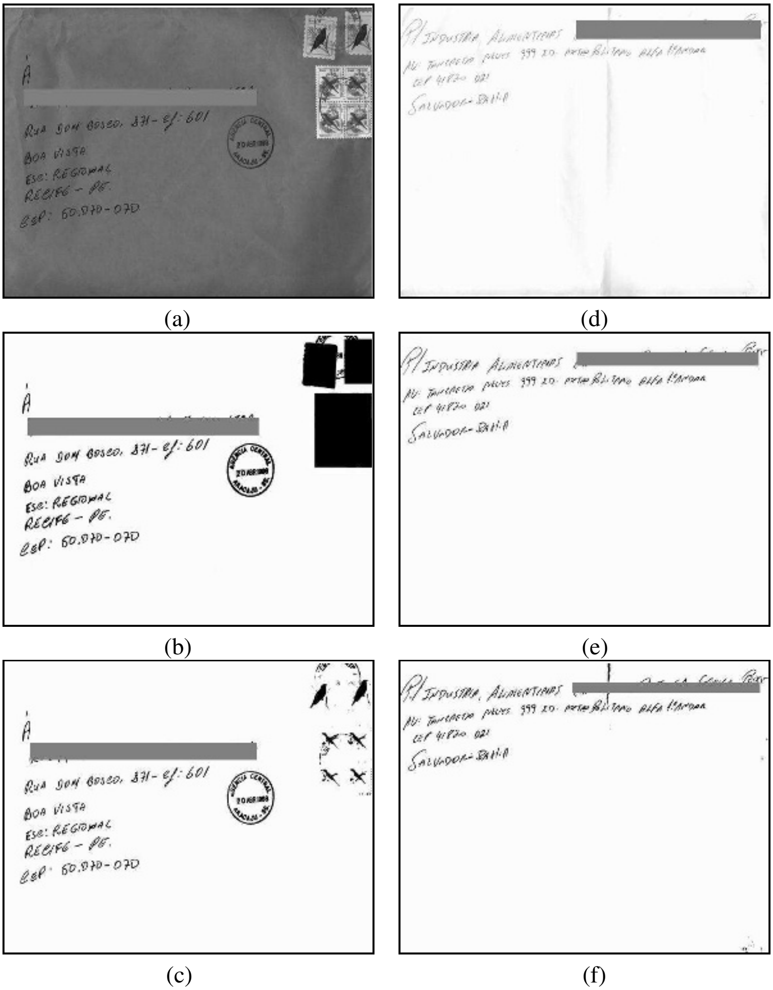
Table 1. Quality evaluation comparison for the database

Method	DBNMSP	DBNMSP	DBSMSP	DBSMSP	$\Upsilon(.)$
	Type I	Type II		II	
Wavelet	0,993	0,645	0,996	0,983	0,404
Fractal	0,917	0,872	0,982	0,967	0,378

Table 2. Quality evaluation comparison for only images No 1 and No 2

Image	DBNMSP	DBNMSP	DBSMSP	DBSMSP	$\Upsilon(.)$
	Type I	Type II		II	
No 1	0,999	0,303	0,999	0,983	0,192
No 2	0,995	0,999	0,997	0,993	0,975

$DBMCP-Type\ II$ measure has shown be more sensitive than the three first ones. The new $\Upsilon(.)$ criterion has shown be much more severe than other ones. This is due to the fact that $\Upsilon(.)$ criterion took in account all the classes. And one can observe that the rates are low because the stamp segmentation was inefficient. The defects occurred in stamp segmentation are similar to bad segmentation simulated from dilation or dilation and salt-pepper noise described in section 3. Figure 5 depicts the segmentation of two postal envelopes, the first one (Figure 5-(a), (b) and (c)) with stamps and the second



(a) (d)

(b) (e)

(c) (f)

Fig. 5. Examples of address block segmentation: (a) Original image No 1, (b) Ground-truth image, (c) Segmentation Result, (d) Original image No 2, (e) Ground-truth image, (f) Segmentation Result

one (Figure 5-(d), (e) and (f)) without stamp neither postmark. Table 2 shows the five discrepancy criteria segmentation rates for the two above images. One can observe that, for Figure 5-(a) where the stamp segmentation was inefficient, whereas $DBMCP - Type\ I$ and $DBSMSP$ and $DBSMSP - II$ rates are high, $DBMCP - Type\ II$ and $\Upsilon(.)$ rates are low. For Figure 5-(d) without stamp neither postmark, the address block segmentation was efficient and $DBMCP - Type\ II$ and $\Upsilon(.)$ rates are very high. Due to noise occurred in address block segmentation, $\Upsilon(.)$ rate is lower than $DBMCP - Type\ II$ one. This point shows that the $\Upsilon(.)$ measure is more able to take in account different segmentation scenarios than other criteria.

5 Conclusions

A new discrepancy evaluation criterion considering different ”scenarios” occurred in a segmentation process have been proposed. The new measure has been compared to traditional discrepancy evaluation criteria. A strategy for evaluating the new measure and other ones in the context of region-based segmentation was used. By applying the discrepancy criteria onto a test database of degradated images, the new discrepancy evaluation criterion has shown to be more sensitive than other ones.

By applying the discrepancy criteria in real segmentation onto wavelet based-segmentation and fractal based-segmentation methods for postal envelope segmentation, experiments have shown that the new measure is more severe than other ones and is able to take in account different segmentation scenarios.

As explained before, evaluation is a critical step. And this study has shown that it is possible to evaluate different segmentation ”scenarios”. In spite of its simplicity, the new measure was shown to be appropriated in the segmentation evaluation challenge. Another advantage is that, in opposite to the study of [5], that excludes bad segmentation, there is no restriction in applying our evaluation approach.

References

1. Zhang, Y.: A survey on evaluation methods for image segmentation. Pattern Recognition **29** (1996) 1335–1346
2. Yasnoff, W., Mui, J.K., Bacus, J.W.: Error measures for scene segmentation. Pattern Recognition **9** (1977) 217–231
3. Menoti, D., Facon, J., Borges, D.L., A.Souza: Segmentation of postal envelopes for address block location: an approach based on feature selection in wavelet space. ICDAR 2003 - 7th InternationalConference on Document Analysis and Recognition **2** (2003) 699–703
4. Eiterer, L.F., Facon, J., Menoti, D.: Fractal-based approach for segmentation of address block in postal envelopes. 9TH Iberoamerican Congress on Pattern Recognition - LNCS Lecture Notes in Computer Science **1** (2004) 454–461
5. Roldán, R.R., Lopera, J.F.G., Allah, C.A., Aroza, J.M., Escamilla, P.L.L.: A measure of quality for evaluating methods of segmentation and edge detection. Pattern Recognition **34** (2001) 969–980