HAL
open science

Comparative study of contour detection evaluation criteria based on dissimilarity measures

Sébastien Chabrier, Hélène Laurent, Christophe Rosenberger, Bruno Emile

To cite this version:

Sébastien Chabrier, Hélène Laurent, Christophe Rosenberger, Bruno Emile. Comparative study of contour detection evaluation criteria based on dissimilarity measures. EURASIP Journal on Image and Video Processing, 2008, pp.1-10. hal-00255983

HAL Id: hal-00255983

https://hal.science/hal-00255983v1

Submitted on 14 Feb 2008

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Comparative study of contour detection evaluation criteria based on dissimilarity measures

Sébastien Chabrier[1], Hélène Laurent[2], Christophe Rosenberger[3] and Bruno Emile[2]

[1] Laboratoire Terre-Océan
Université de la Polynésie française
B.P. 6570
98702 FAA'A, Tahiti - Polynésie Française

[2] Institut PRISME
ENSI de Bourges - Université d'Orléans
88 boulevard Lahitolle
18020 BOURGES Cedex, France

[3] Laboratoire GREYC
ENSICAEN - Université de Caen - CNRS
6 boulevard du Marchal Juin
14050 CAEN Cedex, France

***Abstract*— We present in this article a comparative study of well-known supervised evaluation criteria that enable the quantification of the quality of contour detection algorithms. The tested criteria are often used or combined in the literature to create new ones. Though these criteria are classical ones, none comparison has been made, on a large amount of data, to understand their relative behaviors. The objective of this article is to overcome this lack using large test databases both in a synthetic an a real context allowing a comparison in various situations and application fields and consequently to start a general comparison which could be extended by any person interested in this topic.**

After a review of the most common criteria used for the quantification of the quality of contour detection algorithms, their respective performances are presented using synthetic segmentation results in order to show their performance relevance face to under-segmentation, over-segmentation or situations combining these two perturbations. These criteria are then tested on natural images in order to process the diversity of the possible encountered situations. The used databases and the following study can constitute the ground works for any researcher who wants to confront a new criterion face to well-known ones.

***Index Terms*— segmentation evaluation, contour detection, supervised evaluation.**

I. Introduction

ONE of the first steps in image analysis consists in image segmentation. This stage, which requires homogeneity or dissimilarity notions, leads to two main approaches based respectively on region or contour detection. The purpose is to group together pixels or to delimit areas that have close characteristics and thus to partition the image into similar component parts. Many segmentation methods based on these two approaches have been proposed in the literature [1]–[3] and this subject still remains a prolific one if we consider the quantity of recent publications in this topic. Nobody has already completely mastered such a step. Depending on the acquisition conditions, the applied basic image processing techniques (such as contrast enhancement, noise removal, etc) and the aimed interpretation objectives, different approaches can be efficient. Each of the proposed methods lays the emphasis on different properties and therefore reveals itself more or less suited to a considered application. This variety often makes it difficult to evaluate the efficiency of a proposed method and places the user in a tricky position because no method reveals itself as being optimal in all cases.

That is the reason why many works have been recently performed to solve the crucial problem of the evaluation of image segmentation results [4]–[10]. The proposed evaluation criteria can be split into two major groups. The first one gathers the so called unsupervised evaluation criteria which consist in the computation of different statistics upon the segmentation result to quantify its quality [11]–[13]. These methods are based on the calculation of numerical values from some chosen characteristics attached to each pixel or group of pixels. These methods have the major advantage of being easily computable without requiring any expert assessment. Nevertheless, most of them are not very robust while using textured images and can also present some important shift if the evaluation criterion and the tested segmentation method are both based on the same statistical measure. In such a case, the criterion will not be able to invalidate some erroneous behaviors of the tested segmentation method. The second group is composed of supervised evaluation criteria which are computed from a dissimilarity measure between a segmentation result and a ground truth of the same image. This reference can either be obtained according to an expert judgement or set during the generation of a synthetic test database: in the case of evaluating contour detection algorithms, the ground truth can either correspond to a manually made contour extraction or, if synthetic images are used, to the contour map from which the data set is automatically computed. Even if these methods inherently depend on the confidence in the ground truth, they are widely used for real applications and particularly for medical ones [14]–[16]. In such a case, the ability of a segmentation method to favor a subsequent interpretation and understanding of the image is taken into account.

We focus in this communication on evaluation criteria dedicated to the contour approach and based on the computation of dissimilarity measures between a segmentation result and a reference contour map constituting the ground truth. All the criteria presented in this study do not therefore require the continuity of the contours. For that reason, they are particularly adapted for the evaluation of the usual first step of background/foreground segmentation algorithms which are commonly composed of a preliminary contour detection algorithm followed by some edge closing method. But they are also essential when applications requiring segments detection and not closed contours are pursued. It can for example concern the detection of rivers or roads in aerial images or the detection of veins in palms images for biometric applications. Until now, none comparative study of classical evaluation criteria

has been made on a large amount of data. Generally, when a new evaluation criterion is proposed, its performances are either tested on a few examples (four or five different images) or on several images corresponding to a single application. Moreover, the performance study is rarely completed by the use of synthetic images. However, a preliminary study in a synthetic context can be very useful to test the behaviors of the evaluation criteria face to often encountered situations like under-segmentation, over-segmentation affecting the contour, presence of noise, etc. Working in a controlled environment often allows to more precisely understand the way how a criterion evolves in some specific situations. We try in this article to overcome this lack using large test databases both in a synthetic and a real context allowing a comparison of classical evaluation criteria in various situations and application fields. These databases and the following study could be the ground works for any researcher who wants to confront a new criterion face to well-known ones.

After a first part devoted to a review of evaluation metrics dedicated to contour segmentation and based on dissimilarity measures, several classical criteria are compared. We first tested the evaluation criteria on synthetic segmentation results we created. We also tested them on three hundred images extracted from the ©Corel database which contains various real images corresponding to different application fields such as medicine, aerial photography, landscape images, etc and corresponding experts contour segmentations [4]. The conducted study shows how these databases can be useful to compare the performances of several criteria and put into obviousness their specific behaviors. Finally, we conclude this study and give different perspectives of works in this topic.

II. Supervised evaluation criteria for contour segmentation methods

The different methods presented in this section can either be applied with synthetic or experts ground truths. In the case of synthetic images, the ground truths are of course totally reliable and have an extreme precision, but are not always realistic. For real applications, the expert ground truth is subjective and the confidence attached to this reference segmentation has to be known. Figure 1 presents the supervised evaluation procedure on a real image extracted from the ©Corel database [4].

The next paragraphs present a review of some classical available metrics used in this supervised context for contour segmentation methods. These criteria have often been the basis for the proposal of new ones, either by being modified or combined.

Let I_{ref} be the reference contours corresponding to a ground truth, I_C the detected contours obtained through a segmentation result of an image I.

A. Detection errors

Different criteria have initially been proposed to measure detection errors [17], [18]. Most of them are based on the following expressions or on various definitions issued from them.

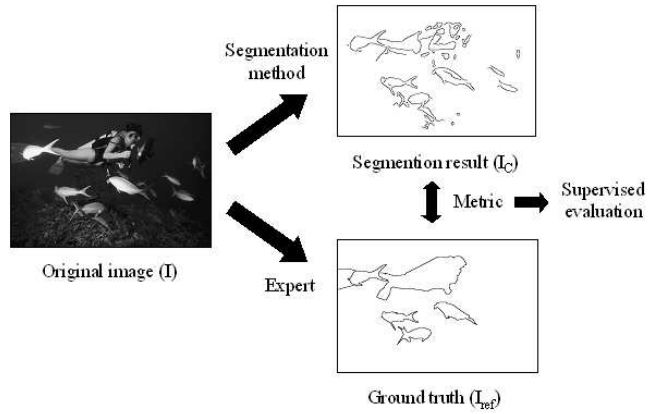

Fig. 1. Supervised evaluation of a segmentation result

The over-detection error (ODE) corresponds to detected contours of I_C which do not match I_{ref}:

$$ODE(I_C, I_{ref}) = \frac{card(I_{C/ref})}{card(I) - card(I_{ref})} \quad (1)$$

where $card(I)$ is the number of pixels of I, $card(I_{ref})$ the number of contour pixels of I_{ref} and $I_{C/ref}$ corresponds to the pixels belonging to I_C but not to I_{ref}.
The under-detection error (UDE) corresponds to I_{ref} pixels which have not been detected:

$$UDE(I_C, I_{ref}) = \frac{card(I_{ref/C})}{card(I_{ref})} \quad (2)$$

where $I_{ref/C}$ corresponds to the pixels belonging to I_{ref} but not to I_C.
Last, the localization error (LE) takes into account the percentage of non-overlapping contour pixels:

$$LE(I_C, I_{ref}) = \frac{card((I_{ref/C}) \cup (I_{C/ref}))}{card(I)} \quad (3)$$

A good segmentation result should simultaneously minimize these three types of error.

Extensions of these detections errors have also been proposed combining them with an additional term taking into account the distance to the correct pixel position [7].

B. Lq and divergence distances

Another idea to compare two images I_C and I_{ref} is to compute between the two images some distance measures [19], [20]. A well-known set of such distances is constituted by the L_q distances:

$$L_q(I_C, I_{ref}) = \left[\frac{\sum_{x \in X} |I_C(x) - I_{ref}(x)|^q}{card(X)} \right]^{\frac{1}{q}} \quad (4)$$

where $I_i(x)$ is the intensity of pixel x in image I_i and with $q \geq 1$. X corresponds to the common domain of I_C and I_{ref}; in our case, X is the complete image. These distances which are initially defined to deal with the intensities of the pixels can also be used for binary images. Note that, among these distances, the classical Root Mean Squared error (RMS) can

be obtained with $q = 2$. For the comparative study, q has been chosen in $\{1, 2, 3, 4\}$ defining the L_1, L_2, L_3 and L_4 distances.

The considered measures can be completed by different distances issued from probabilistic interpretations of images: the Küllback and Bhattacharyya (DKU and DBH) distances and the "Jensen-like" divergence measure (DJE) based on Rènyi entropies [21].

$$DKU(I_C, I_{ref}) = \frac{\sum_{x \in X} \big(I_C(x) - I_{ref}(x)\big) \times Log\left(\frac{I_C(x)}{I_{ref}(x)}\right)}{card(X)} \tag{5}$$

$$DBH(I_C, I_{ref}) = -Log\left(\frac{\sum_{x \in X} \sqrt{I_C(x) \times I_{ref}(x)}}{card(X)}\right) \tag{6}$$

$$DJE(I_C, I_{ref}) = J_1\left(\frac{I_C(x) + I_{ref}(x)}{2}, I_C(x)\right) \tag{7}$$

with

$$J_1\big(I_C(x), I_{ref}(x)\big) = H_\alpha\left(\sqrt{I_C(x) \times I_{ref}(x)}\right) - \frac{H_\alpha\big(I_C(x)\big) + H_\alpha\big(I_{ref}(x)\big)}{2} \tag{8}$$

where H_α corresponds to the Rènyi entropies parametrized by $\alpha > 0$. This parameter is set to 3 in the comparative study [22].

If these measures permit to obtain a global comparison between two images, they are often described in the literature as not correctly transcribing the human visual perception and more particularly the topological transformations (translations, rotations, etc). The concerned gray-level domain is indeed not taken into account. If gray-level images are used, a same intensity difference will then be equally penalized whatever the domain can be. In our case, these distances are used with binary images, this drawback does therefore not exist anymore. In the same way, the global position information does not intervene in distance computation. Thus, if the same object appears in the two images with a simple translation, the distances will increase in an important way. If this evolution can be disturbing with an object detection objective for example, it becomes an advantage in our case where a contour translation is a mistake.

C. Hausdorff distance

The Hausdorff distance between two pixels sets is computed as follows [23]:

$$HAU(I_C, I_{ref}) = max\Big(h(I_C, I_{ref}), h(I_{ref}, I_C)\Big) \tag{9}$$

where

$$h(I_C, I_{ref}) = \max_{a \in I_C}\left(\min_{b \in I_{ref}} \| a - b \|\right) \tag{10}$$

$HAU(I_C, I_{ref}) = d$ then means that all the pixels belonging to I_C are not farther than d from some pixels of I_{ref}. Although this measure is theoretically very interesting and can give a good similarity measure between the two images, it is described as being very noise sensitive.

Several extensions of this measure, like the Baddeley distance, can be found in the literature [24].

D. Pratt's figure of merit

This criterion [25] corresponds to an empirical distance between the ground truth contours I_{ref} and those obtained with the chosen segmentation I_C:

$$PRA(I_{ref}, I_C) = \frac{1}{max\{card(I_{ref}), card(I_C)\}} \sum_{k=1}^{card(I_C)} \frac{1}{1 + d^2(k)} \tag{11}$$

where $d(k)$ is the distance between the kth pixel belonging to the segmented contour I_C and the nearest pixel of the reference contour I_{ref}.

This measure has no theoretical proof but is however one of the most used descriptors. It is not symmetrical and does not express under-segmentation or shape errors. Moreover, it is also described as being sensitive to over-segmentation and localization problems. To illustrate some limits of this criterion, we present in figure 2 different situations with an identical number of misclassified pixels and leading to the same criterion value.

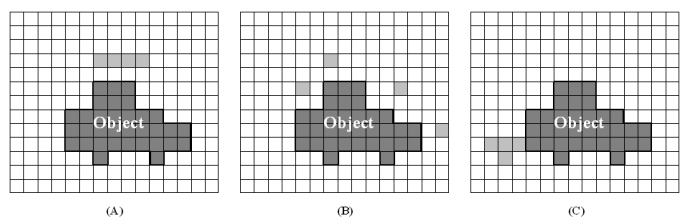

Fig. 2. Different situations with an identical number of misclassified pixels and leading to the same criterion value

The three depicted situations are very dissimilar and should not be equally marked. The misclassified pixels should belong to the object in figure 2(C) and to the background in figure 2(A). The proposed criterion considers these situations as equivalent although the consequences on the object size and shape are totally different. Moreover, this criterion does not discriminate between isolated misclassified pixels (figure 2(B)) or a group of such pixels (figure 2(A)) though the last situation is more prejudicial.

Modified versions of this criterion have been proposed in the literature [26].

E. Odet's criteria

Different measurements have been proposed in [27] to estimate various errors in binary segmentation results. Amongst them, two divergence measures seem to be particularly interesting. The first one (OCO) evaluates the divergence between the over-segmented contour pixels and the reference contour pixels:

$$OCO(I_C, I_{ref}) = \frac{1}{N_o} \sum_{k=1}^{N_o} \left(\frac{d(k)}{d_{TH}}\right)^n \tag{12}$$

where $d(k)$ is the distance between the kth pixel belonging to the segmented contour I_C and the nearest pixel of the

reference contour I_{ref}. N_o corresponds to the number of over-segmented pixels. d_{TH} is the maximum distance, starting from the segmentation result pixels, allowed to search for a contour point. If a pixel of the segmentation result is farther than d_{TH} from the reference, the criterion value is highly penalized (all the more since n is big), the quotient $d(k)/d_{TH}$ exceeding one. n is a scale factor which permits to weight the pixels depending on their distance from the reference contour.

The second one (OCU) estimates the divergence between the under-segmented contour pixels and the computed contour pixels:

$$OCU(I_C, I_{ref}) = \frac{1}{N_u} \sum_{k=1}^{N_u} \left(\frac{d_u(k)}{d_{TH}} \right)^n \tag{13}$$

where $d_u(k)$ is the distance between the kth non-detected pixel and the nearest one belonging to the segmented contour. N_u corresponds to the number of under-segmented pixels.

These two criteria take into account the relative position for the over- and under-segmented pixels. The threshold d_{TH}, which has to be set according to each application precision requirement, permits to take the pixels into account differently with regard to their distance from the reference contour. These criteria also allow, thanks to exponent n, to differently weight the estimated contour pixels that are close to the reference contour and those whose distance to the reference contour is close to d_{TH}. With a small value of n, the first ones are privileged, which leads to a precise evaluation. For the comparative study, n is set to 1 and d_{TH} equals 5.

F. Discussion

As previously exposed, most of the presented criteria are based on the computation of distance measures between a segmentation result and a ground truth. Even if the principles are often quite similar, no comparison has been realized in the literature to evaluate the relative performances of these proposed criteria. The problem lies in the fact that the reference is not always easily available. Though a few databases of assessed real images exist, a preliminary study on synthetic images seems to be a powerful manner to make a reliable comparison. Working in a controlled environment indeed allows to more precisely understand the way how a criterion evolves in some specific situations like under-segmentation, over-segmentation affecting the contour, presence of noise, etc.

III. Comparative study

When new evaluation criteria are proposed in the literature, the definitions and principles on which they are based are of course exposed. Thereafter, their behaviors are generally illustrated by a few examples, often on some segmentation results of a chosen image. A comparative study with classical existing methods is sometimes conducted on a limited test database. However, a comparative study of the principal evaluation criteria, made on a large amount of data and enabling to determine their relative relevance and their favored application contexts, is not systematically done. We try to fill this lack in this section. The main supervised evaluation criteria defined for contour segmentation results and previously exposed are here tested. They mainly rely on the computation of distances between an obtained segmentation result and a ground truth. The tested criteria are ODE, UDE, LE, L_1, L_2, L_3, L_4, DKU, DBH, DJE, HAU, PRA, OCO and OCU. In order to make the comparison easier for the reader, we made all the criteria evolve in the same way. They all are positive, growing with the amplitude of the perturbations. The value 0 corresponds therefore to the best result. We first studied the criteria on synthetic segmentation results. Afterwards, we tested the chosen criteria on a selection of real images extracted from the ©Corel database for which manual segmentation results provided by experts are available [4]. Contrary to synthetic cases, this database allows us to process the diversity of the possible encountered situations in natural images. Indeed, it contains images corresponding to different application fields such as aerial photography or landscape images.

A. Preliminary study on synthetic segmentation results

In order to study the behaviors of the previously presented criteria in the face of different perturbations, we first generated some synthetic segmentation results corresponding to several degradations of a ground truth we created. Some of the obtained results were described in [28]; we present in this article the complete study.

The used ground truth is composed of five components: a central ring and four external contours (see figure 3). The tested perturbations are the following:

- under-segmentation: one or several components of the ground truth are missing,
- over-segmentation affecting the complete image: noisy ground truth with impulsive noise (probability from 0.1% to 50%),
- over-segmentation affecting the contour area: from 1 to 5 dilatation processes,
- over- and under-segmentation affecting the contour area: impulsive noise (probability of 1%, 5%, 10% or 25%) in the contour area (width from 1 to 5 pixels),
- localization error: synthetic segmentation results obtained by contour shifts from 1 to 5 pixels in the four cardinal directions.

Different examples of the considered perturbations are presented in figure 3.

Figure 4 presents the evolution of four criteria (L_1, HAU, OCO, OCU) in the face of under-segmentation. The $Y-$coordinates of the curves present the criteria values, the $X-$coordinates correspond to the different segmentation results to assess. Four of them (results 4, 11, 15 and 28) are presented in figure 4 and are put into obviousness on the curves thanks to bold or dotted lines. OCO is equal to zero whatever case is considered. As OCO only measures over-segmentation, it equivalently grades a segmentation result with one or several components missing. ODE has the same behavior. L_1 presents different stages allowing to gradually penalize under-segmentation. This behavior corresponds to the expected one and the majority of the criteria evolves in that way (UDE, LE, L_1, L_2, L_3, L_4, DKU, DBH, DJE,

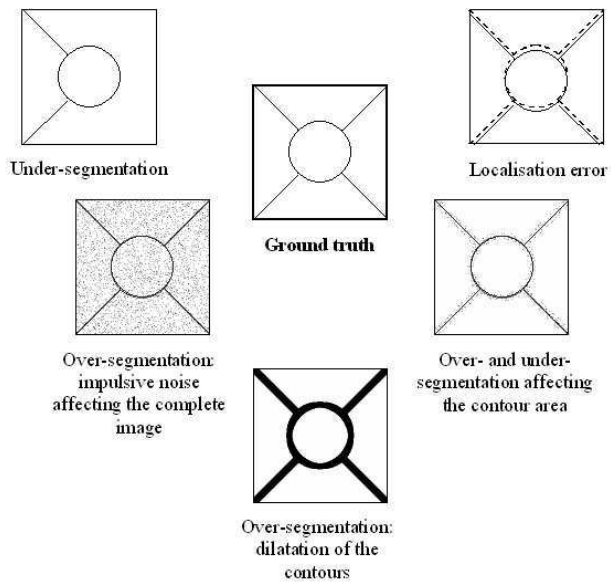

Fig. 3. Ground truth and examples of perturbations

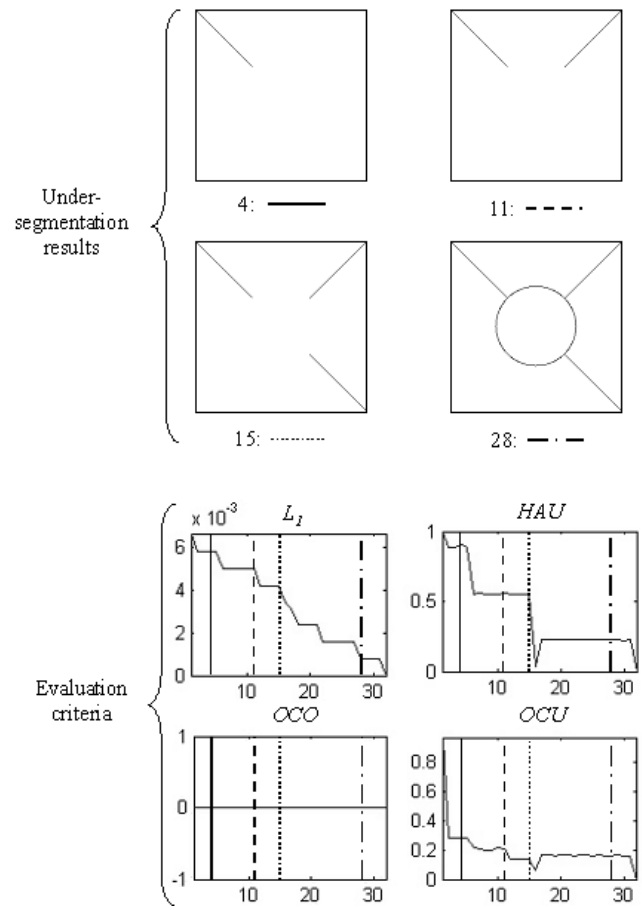

Fig. 4. Evolution of four evaluation criteria in the face of under-segmentation

PRA). HAU also presents a graduated evolution but seems to suffer from a lack of precision. It equivalently grades some segmentation results even if the number of detected components is completely different (see for example the segmentation results 11 and 15). Finally, OCU, which normally measures under-segmentation, does not allow to correctly differentiate the synthetic segmentation results. For example, it better grades result 15 than result 28.

Figure 5 presents the evolution of three criteria (DKU, PRA, OCO) in the face of over-segmentation corresponding to the presence of impulsive noise. OCO penalizes too strongly the presence of over-segmentation: for example, it equivalently grades the segmentation results with impulsive noise of probability 0.2% and 25%. Moreover, the evolution of this criterion is not monotonic. HAU has the same kind of behavior. DKU really penalizes over-segmentation only when it reaches a high level. ODE, LE, L_1, L_2, L_3, L_4, DBH, DJE have the same kind of behavior. OCU and UDE, that only measure under-segmentation, equivalently grade segmentation results with a small or high presence of noise. They are equal to zero whatever case is considered. Finally, PRA permits to penalize the presence of impulsive noise as soon as it appears. This criterion is the only one with a behavior that is close to the human decision: an expert will notice the presence of noise even for a small proportion and will immediately penalize it. On the other hand, an expert will not grade too noisy segmentation results very differently.

Concerning over-segmentation due to the dilatation of contours, except UDE and OCU which are equal to zero whatever case is considered, the other criteria present quite the same behavior which is the expected one: figure 6 presents as an example the evolution of LE and L_2.

In order to test the influence of combined over- and under-segmentation, we first added, in the contour area, an impulsive noise with a probability of 1%, 5%, 10% and 25%. The noise was respectively added in a neighborhood of the contour with a window width from 1 to 5 pixels. Figure 7 presents the evolution of three criteria (DJE, HAU, PRA) in the face of this perturbation. We can notice that, as expected, HAU ranks the segmentation results with respect to the width of the noisy area around the contour. Nevertheless, it does not seem to take into account the probability of apparition of noise: the three examples presented in figure 7 are equivalently graded. HAU and OCO, that evolve in the same way, seem to suffer from a lack of precision in that case. On the other hand, DJE and PRA correctly evolve penalizing in a more important way a high probability and a large noisy area around the contour. Most of the other criteria: LE, ODE, DBH, DKU, L_1, L_2, L_3 and L_4 have the same behavior.

Last, we studied the influence of localization error. For these synthetic segmentation results, the contours have been moved from 1 to 5 pixels in the four cardinal directions. Figure 8 presents the evolution of three criteria (ODE, UDE, PRA) in the face of this perturbation. In this figure, the original contour appears dotted to make the perturbation remarkable. We can observe that all the criteria penalize more a segmentation result if it corresponds to an increasing shifting. Whatever, UDE and PRA are more precise (OCO, OCU and HAU evolve in a similar way).

As a result of this preliminary study, we can conclude that most of the studied criteria have a global correct be-

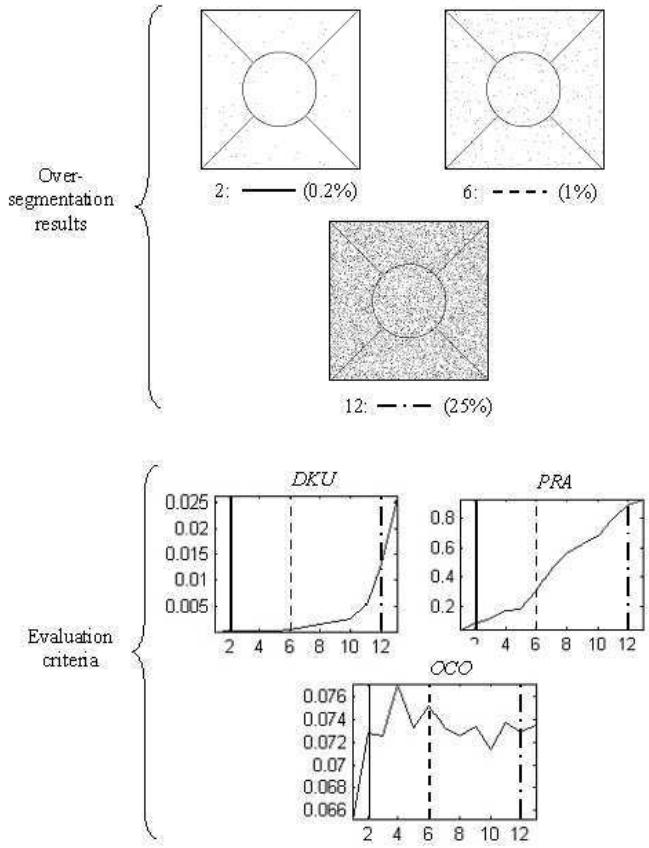

Fig. 5. Evolution of three evaluation criteria in the face of over-segmentation corresponding to the presence of impulsive noise

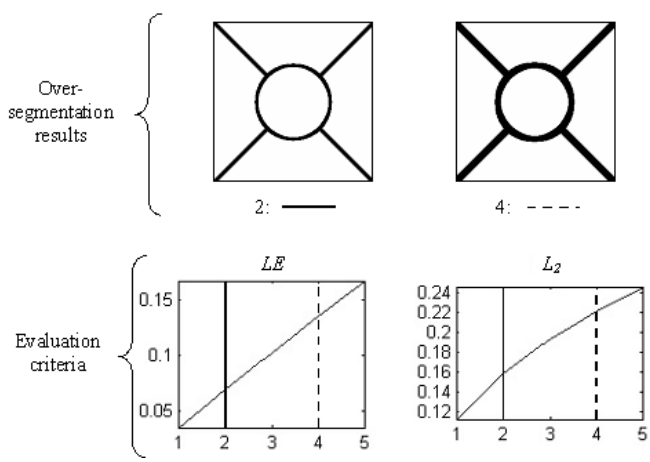

Fig. 6. Evolution of two evaluation criteria in the face of over-segmentation due to the dilatation of contours

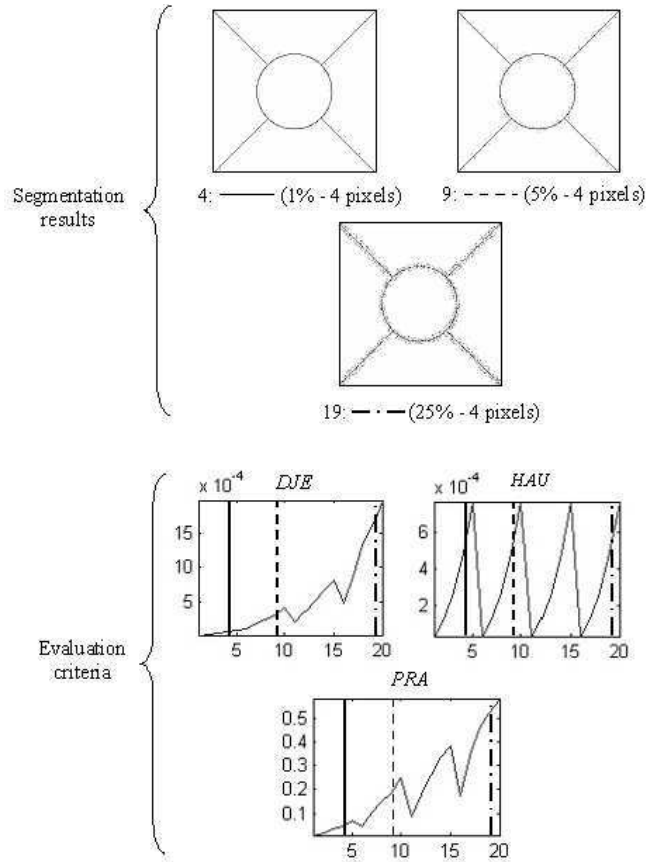

Fig. 7. Evolution of three evaluation criteria in the face of combined over-and under-segmentation localized in the contour area

	Under-segmentation	Over-segmentation		Over-/under-segmentation	Localization error
		Noise	Dilatation		
ODE		**	***	***	**
UDE	***				***
LE	***	**	***	***	**
L_1	**	**	***	***	**
L_2	**	**	***	***	**
L_3	**	**	***	***	**
L_4	**	**	***	***	**
DKU	***	**	***	***	**
DBH	***	**	***	***	**
DJE	***	**	***	***	**
HAU	*	*	***	*	***
PRA	***	***	***	***	***
OCO		*	***	*	***
OCU	*				***

TABLE I

RELEVANCE OF THE DIFFERENT CRITERIA FOR EACH CONSIDERED PERTURBATION (THE MORE STARS, THE BETTER CRITERION).

havior, that is a behavior corresponding in general to the expected one. However, some of them turned out not to be appropriate to characterize some situations. Table I sums up the performances of the different criteria in the face of the considered perturbations. The OCO and OCU criteria were computed with the parameters advocated in [27] (n=1 and d_{TH}=5). Fitted parameters seem to be essential to obtain the optimal performances for each situation. This shows that these criteria are less generic than ODE or UDE. These conclusions could be useful to make the necessary choices to propose a new measure combining two criteria dedicated respectively to under- and over-segmentation. HAU revealed itself as being not relevant to precisely characterize under-segmentation or localization errors. Finally, LE, L_1, L_2, L_3, L_4, DKU, DBH, DJE and PRA have a correct behavior in the face of the considered perturbations, PRA giving in this preliminary study the most clear-cut decision.

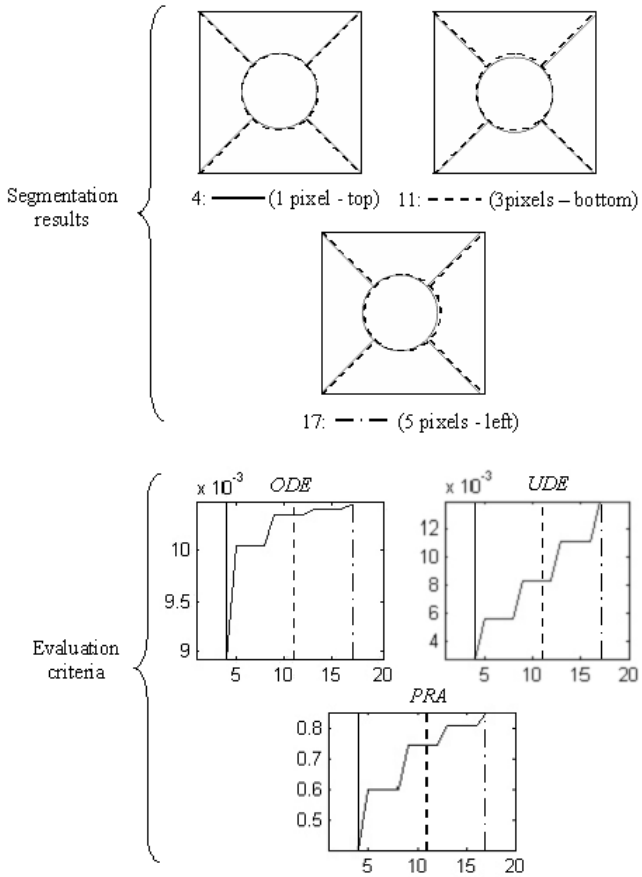

Fig. 8. Evolution of three evaluation criteria in the face of combined over-and under-segmentation due to contours shifting

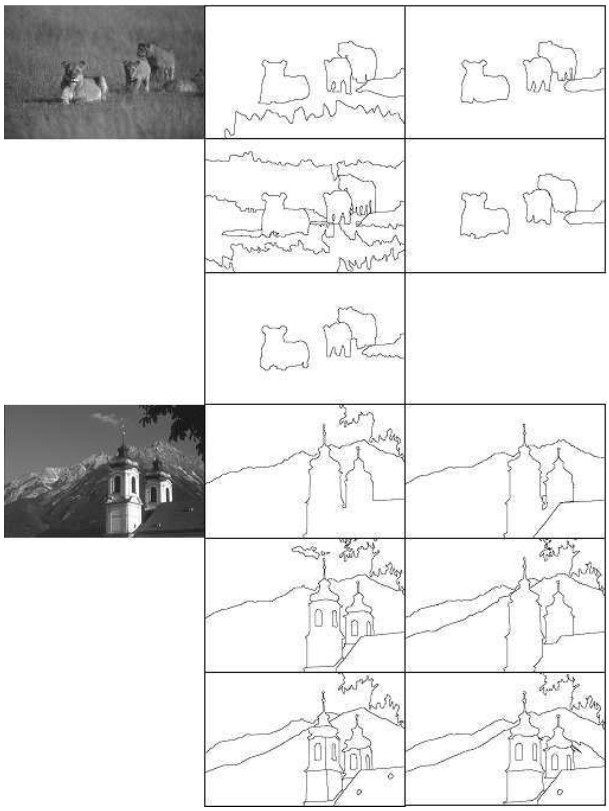

Fig. 9. Examples of real images extracted from the ©Corel database and corresponding experts ground truths

B. Complementary study on real segmentation results

In order to complete this preliminary study, we tested the different criteria on segmentation results issued from real images to process the diversity of the possible encountered situations. Our database was composed of 300 images extracted from the ©Corel database for which manual segmentation results provided by experts are available [4]. Figure 9 presents two examples of the available images and corresponding ground truths established by different experts. For each image of the database, 5 to 8 experts ground truths are available.

We can notice that these ground truths can be quite dissimilar. Some experts only attach to put into obviousness the main objects in the image. Others are more sensitive to the objects present in the background. We then decided to make a fusion of the different expert ground truths in order to obtain a more representative one. The following method was applied to create the fused ground truths: for each expert ground truth, a widened one was created. The pixels belonging to the contour were set to 3, their direct neighbors (4-connected) were set to 2 and the following ones, connected to direct neighbors, were set to 1. For one real image, all the available widened ground truths were added and a pixel was considered as belonging to the contour if its score strictly exceeded twice the number of experts. Figure 10 presents the principle on which the fused ground truths were established and figure 11 presents the fused ground truths obtained for two real images.

In order to test the different evaluation criteria, we segmented the image database with 10 segmentation algorithms based on threshold selection [29]:

- color gradient,
- texture gradient,
- second moment matrix,
- brightness/texture gradients,
- gradient multi-scale magnitude,
- brightness gradient,
- first moment matrix,
- color/texture gradients,
- gradient magnitude,
- Canny filter.

These filters generate fuzzy contour maps. Figure 12 presents examples of the maps obtained for two images with the Canny filter.

As we need binary contour maps, we thresholded the fuzzy contour maps to obtain various segmentation results. The threshold value (Th) was set from 5 to 255. For each segmentation result, the 14 studied criteria were computed using the fused ground truth. Figures 13 and 14 present the different curves obtained with the Canny filter on two images of the ©Corel database. The Y-coordinates of the curves present the criteria values. The X-coordinates correspond to the different chosen values ($Th \in [5, 255]$) to threshold the fuzzy contour map; a very small threshold value conducting to a high over-segmented segmentation result. In order to make the comparison easier for the reader, we normalized the criteria: they all evolve between 0 and 1, 0 being the best result.

A relevant criterion should be able to detect a compromise between under- and over-segmentation and consequently

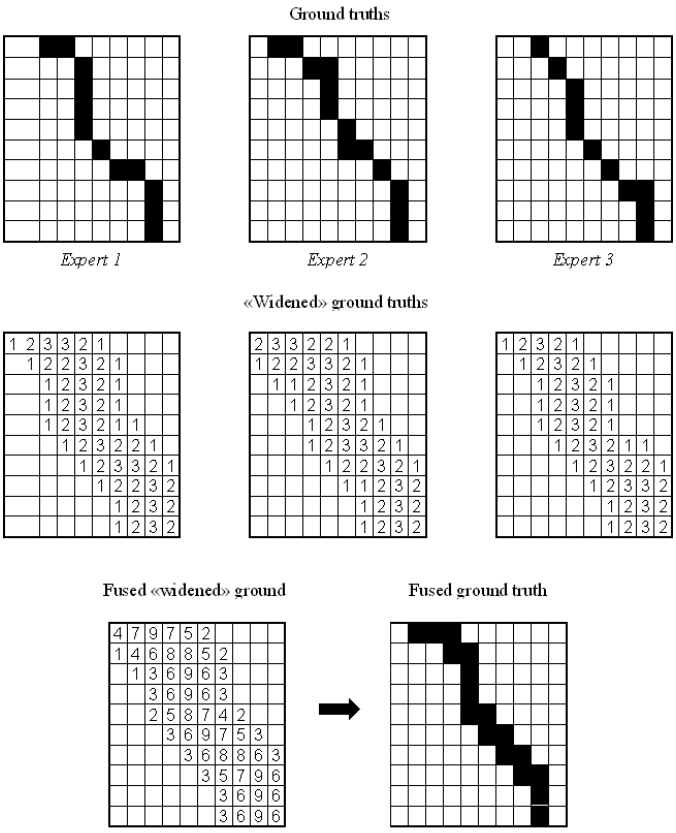

Fig. 10. Principle on which the fused ground truths are created

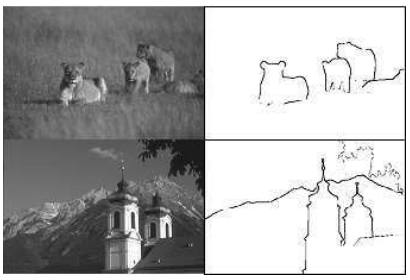

Fig. 11. Examples of obtained fused ground truths

Fig. 12. Example of the fuzzy contour map obtained for two original images of the ©Corel database with the Canny filter

present a minimum. This approach is similar to the one proposed in [7]. A criterion which evolves in a monotonic way is indeed not satisfactory. If it always increases (respectively decreases) that means that the over-segmented (resp. the under-segmented) case is too much favored. Similarly, even it is not monotonic, a criterion which systematically selects the first tested threshold value: $Th = 5$ (resp. the last tested threshold value: $Th = 255$) as being the best, must be rejected.

Figures 13 and 14 present, for two images of the ©Corel database, the evolution of the 14 studied criteria for segmentation results obtained with the Canny filter using different thresholds.

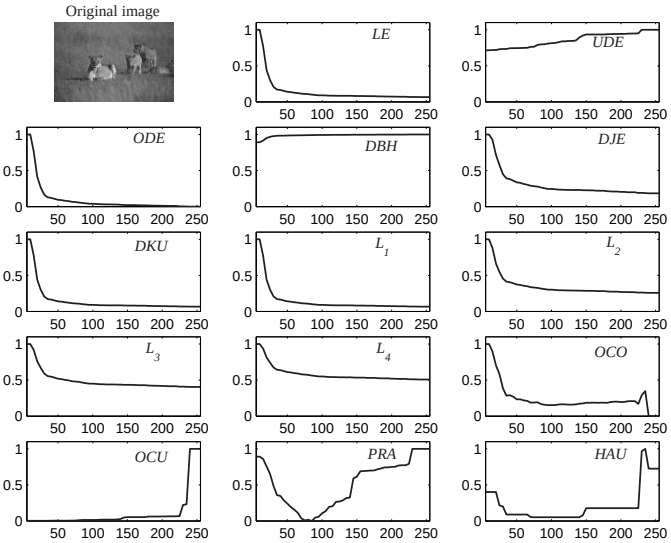

Fig. 13. Evolution, for one image of the ©Corel database, of the 14 studied criteria for segmentation results obtained with the Canny filter using different thresholds.

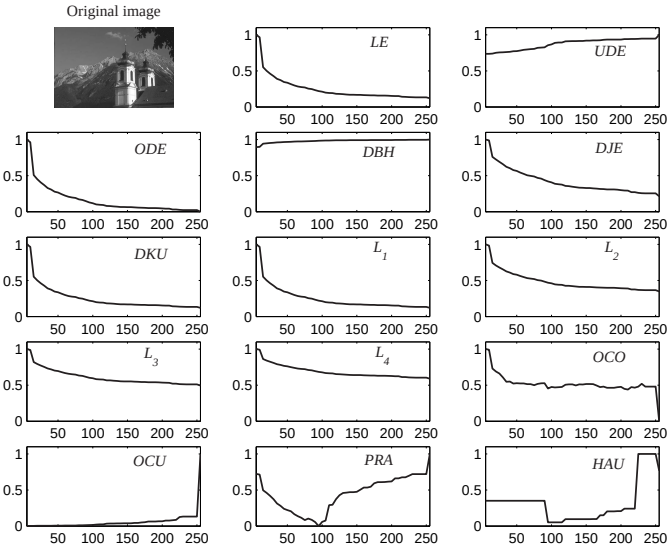

Fig. 14. Evolution, for one image of the ©Corel database, of the 14 studied criteria for segmentation results obtained with the Canny filter using different thresholds.

We can observe, on both figures 13 and 14, that the LE, L_1, L_2, L_3, L_4, DJE, DKU criteria are always decreasing, preferring the under-segmentation. As a result of their definitions, OCO and ODE also privilege the under-segmentation. Similarly UDE and OCU privilege the over-segmentation. We can also notice that DBH is not relevant. First of all, it evolves in a monotonic way and the obtained values are

very similar whatever case is considered, high over- or under-segmentation. These results allow to balance the conclusions resulting from the preliminary study using synthetic segmentation results. It shows the interest to complete the study with real segmentation results. Finally, only two criteria allow to detect a compromise: PRA and HAU. We can however notice, as previously mentioned in the preliminary study on synthetic segmentation results, that HAU seems to suffer from a lack of precision. It equivalently grades some segmentation results even if a different threshold value always conducts to slightly different situations (see for example figure 14: for a threshold value growing from 5 to 90, HAU is constant).

Figure 15 presents the binary images obtained using the optimal threshold selected by the criterion PRA for the two original images of figures 13 and 14 with the Canny filter.

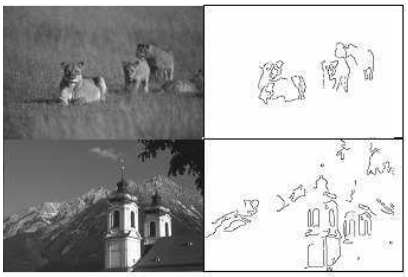

Fig. 15. Binary images obtained using the optimal threshold selected by the criterion PRA for the two original images of figures 13 and 14 with the Canny filter.

Figure 16 presents the mean curves obtained on the 300 images of the ©Corel database using for each image the 10 segmentation algorithms. If these curves only present the global trends of the criteria behaviors, they are nevertheless revealing. Some of them are very similar with those presented in the single cases of figures 13 and 14 expressing repetitive behaviors. The two criteria presenting a minimum are PRA and HAU. These two criteria allow in almost all cases to detect a compromise.

Table II sums up the situation mostly favored by the different criteria in the face of segmentation results issued from real images of the ©Corel database.

	Under-segmentation	Compromise	Over-segmentation
ODE	√		
UDE			√
LE	√		
L_1	√		
L_2	√		
L_3	√		
L_4	√		
DKU	√		
DBH			√
DJE	√		
HAU		√	
PRA		√	
OCO	√		
OCU			√

TABLE II

SITUATION MOSTLY FAVORED BY THE CRITERIA FOR SEGMENTATION RESULTS ISSUED FROM REAL IMAGES OF THE ©COREL DATABASE.

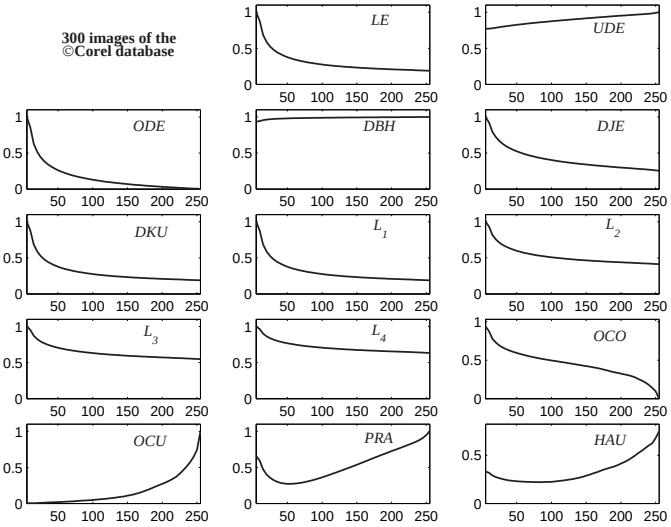

Fig. 16. Mean evolution, for the 300 images of the ©Corel database, of the 14 studied criteria for segmentation results obtained with 10 segmentation algorithms based on threshold selection.

IV. CONCLUSION

We presented in this article a review of classical available metrics used for the evaluation, in the supervised context, of contour detection methods. The studied criteria compute a dissimilarity measure between a segmentation result and a ground truth. We tested their relative performances on synthetic and real segmentation results. Thanks to the first part of the comparison, done on synthetic results, we concluded that different criteria (LE, L_1, L_2, L_3, L_4, DKU, DBH, DJE and PRA) had a global correct behavior. PRA stood out as the most interesting one, giving more discriminated results and allowing a most clear-cut decision. The second part of the comparative study, done on real segmentation results, confirmed this conclusion.

This article permitted to start a general comparison which could be extended by any person interested in this topic. The used databases are at everyone's disposal at the following addresses:

- http://www.ecole.ensicaen.fr/∼rosenber/ressources.html for the synthetic segmentation results
- http://www.eecs.berkeley.edu/Research/Projects/CS/vision /grouping/segbench/ for the real segmentation results extracted from the ©Corel database.

This study concerned criteria which do not require the continuity of the contours, we plan to first of all complete it using criteria dedicated to the evaluation of region detection algorithms when segmentations presenting closed contours are available (at least closed by the image edges). In these cases, the correspondence between contours and regions can be easily obtained.

Secondly, we plan to combine different criteria in order to obtain a new one taking advantage of their relative specificities. It could be for example interesting to combine OCO and OCU which are respectively dedicated to the detection of over- and under-segmentation.

We are also interested in assessing if a criterion is able to reflect the subjective evaluation of a human expert or not. We plan to realize a psychovisual study for the comparison of contour segmentation results. The goal of this experiment will be first of all to know if the comparison of multiple contour segmentation results of a single image can be made easily and can provide a similar judgement for different experts. This psychovisual study could also be used to check if evaluation criteria are able to reproduce the human judgment.

These evaluation criteria could finally be applied in medical contexts when comparisons with expert diagnostics are required. When new segmentation methods are proposed in this context, their behaviors are often illustrated by few examples and generally visually assessed. An evaluation criterion will permit to overcome this subjective step or to confirm it.

Acknowledgments

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

References

[1] R. H. Haralick and L. G. Shapiro, "Image segmentation techniques," *Computer Vision, Graphics and Image Processing (CVGIP)*, vol. 29, pp. 100–132, 1985.

[2] M. Heath, S. Sarkar, T. Sanocki, and K. Bowyer, "Comparison of edge detectors: A methodology and initial study," *Computer Vision and Image Understanding (CVIU)*, vol. 69, no. 1, pp. 38–54, january 1996.

[3] J. Freixenet, X. Muñoz, D. Raba, J. Marti, and X. Cufi, "Yet another survey on image segmentation: Region and boundary information integration," *European Conference on Computer Vision (ECCV)*, pp. 408–422, 2002.

[4] D. Martin, C. Fowlkes, D. Tal, and J. Malik, "A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics," *International Conference on Computer Vision (ICCV)*, pp. 416–423, July 2001.

[5] G. Liu and R. M. Haralick, "Optimal matching problem in detection and recognition performance evaluation," *Pattern Recognition*, vol. 35, no. 10, pp. 2125–2139, 2002.

[6] Y. Yitzhaky and E. Peli, "A method for objective edge detection evaluation and detector parameter selection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 8, pp. 1027–1033, 2003.

[7] N. L. Fernandez-Garcia, R. Medina-Carnicer, A. Carmona-Poyato, and M. P.-V. F.J. Madrid-Cuevas, "Characterization of empirical discrepancy evaluation measures," *Pattern Recognition Letters*, vol. 25, no. 1, pp. 35–47, 2004.

[8] S. Chabrier, "Contribution à l'évaluation de performances en segmentation d'images," Ph.D. dissertation, Université d'Orléans, 2005.

[9] S. Wang, F. Ge, and T. Liu, "Evaluating edge detection through boundary detection," *EURASIP Journal on Applied Signal Processing*, pp. 1–15, 2006.

[10] Y. Zhang, Ed., *Advances in image and video segmentation*. IRM Press, Hershey, PA, USA, 2006.

[11] R. Zeboudj, "Filtrage, seuillage automatique, contraste et contours: du pré-traitement à l'analyse d'image," Ph.D. dissertation, Université de Saint Etienne, 1988.

[12] M. Borsotti, P. Campadelli, and R. Schettini, "Quantitative evaluation of color image segmentation results," *Pattern Recognition Letters*, vol. 19, pp. 741–747, 1998.

[13] C. Rosenberger, "Mise en oeuvre d'un système adaptatif de segmentation d'images," Ph.D. dissertation, Université de Rennes 1, december 1999.

[14] S. Montresor, M. J. Lado, P. G. Tahoces, M. Souto, and J. J. Vidal, "Analytic wavelets applied for the detection of microcalcifications. a tool for digital mammography," *European Signal Processing Conference (EUSIPCO)*, pp. 2215–2218, 2004.

[15] F. Marques, G. Cuberas, A. Gasull, D. Seron, F. Moreso, and N. Joshi, "Mathematic morphology approach for renal biopsy analysis," *European Signal Processing Conference (EUSIPCO)*, pp. 2195–2198, 2004.

[16] W. W. Lee, I. Richardson, K. Gow, Y. Zhao, and R. Staff, "Hybrid segmentation of the hippocampus in mr images," *European Signal Processing Conference (EUSIPCO)*, 2005.

[17] W. A. Yasnoff, J. K. Mui, and J. W. Bacus, "Error measures for scene segmentation," *Pattern Recognition*, vol. 9, pp. 217–231, 1977.

[18] Y. J. Zhang, "A survey on evaluation methods for image segmentation," *Pattern Recognition*, vol. 29, no. 8, pp. 1335–1346, 1996.

[19] H. Laurent, "Détection de ruptures spectrales dans le plan temps-fréquence," Ph.D. dissertation, Université de Nantes, november 1998.

[20] M. Basseville, "Distance measures for signal processing and pattern recognition," *Signal processing*, vol. 18, no. 4, pp. 349–369, 1989.

[21] O. Michel, R. Baraniuk, and P. Flandrin, "Time-frequency based distance and divergence measures," *Proceedings of the IEEE International Symposium on TF and TS Analysis (TFTS)*, pp. 64-67, Philadelphia, USA 1994.

[22] R. Baraniuk, P. Flandrin, and O. Michel, "Information and complexity on the time-frequency plane," *Proceedings of the 14th Symposium GRETSI*, vol. 1, pp. 359-362, Juan-les-Pins, France 1993.

[23] M. Beauchemin, K. Thomson, and G. Edwards, "On the hausdorff distance used for the evaluation of segmentation results." *Canadian Journal of Remote Sensing (CJRS)*, vol. 24(1), pp. 3–8, 1998.

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

[25] W. Pratt, O. D. Faugeras, and A. Gagalowicz, "Visual discrimination of stochastic texture fields," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 8, no. 11, pp. 796–804, 1978.

[26] K. Strasters and J. J. Gerbrands, "Three-dimensional image segmentation using a split, merge and group approach," *Pattern Recognition Letters*, vol. 12, pp. 307–325, 1991.

[27] C. Odet, B. Belaroussi, and H. Benoit-Cattin, Eds., *Scalable Discrepancy Measures for Segmentation Evaluation*, vol. 1, septembre 2002.

[28] S. Chabrier, H. Laurent, C. Rosenberger, and Y.-J. Zhang, "Supervised evaluation of synthetic and real contour segmentation results," *European Signal Processing Conference (EUSIPCO)*, 2006.

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

Sébastien Chabrier Sebastien Chabrier is assistant professor at the Pacific University (France). He obtained his Phd from the University of Orleans in 2005. He belongs to the Terre-Océan laboratory. His research interest concerns segmentation and its evaluation.

Hélène Laurent Hélène Laurent is assistant professor at ENSI of Bourges (France). She obtained her Phd from the University of Nantes in 1998. She belongs to the PRISME intitute (SEISME research unit). Her research interests concern segmentation evaluation and pattern recognition.

Christophe Rosenberger Christophe Rosenberger is full professor at ENSICAEN (France). He obtained his Phd from the University of Rennes I in 1999. He belongs to the GREYC laboratory (Image research unit). His research interests concern evaluation of image processing and biometrics.

Bruno Emile Bruno Emile is assistant professor at IUT of Châteauroux (France). He obtained his Phd from the University of Nice in 1996. He belongs to the PRISME institute (SEISME research unit). His research interests concern segmentation evaluation and object detection.