Multi-scale criteria for the evaluation of image segmentation algorithms

Sylvie Philipp-Foliguet*, Laurent Guigues**
*ETIS, UMR CNRS 8051
6 avenue du Ponceau, 95014 Cergy Cedex
**CREATIS, UMR CNRS 5220, Inserm U 630
INSA-Lyon, 7 Avenue Jean Capelle, F-69621 Villeurbanne Cedex
Email: philipp@ensea.fr, laurent.guigues@creatis.insa-lyon.fr

***Abstract*— This paper deals with evaluation of image segmentation methods. We start with a state-of-the art of the evaluation criteria, involving a reference segmentation or not. Based on an analysis of the main existing criteria, we propose new criteria, when no ground-truth is available. These criteria, based on an energetic formalism, take into account both the complexity of the segmented image, through the boundary length and the goodness-of-fit of an underlying model with the initial data. The main interest is not to fix the level of detail expected by the evaluation criterion but to leave it to the user's choice, according to his purpose. These evaluation criteria are thus multi-scale criteria. Various forms of energy formulation are experimentally compared. Then after having chosen a particular energy, the performances of this new criterion is compared to the main existing criteria.**

***Index Terms*— Image, evaluation, segmentation, criterion, energy, scale**

I. INTRODUCTION

Because of the profusion of image segmentation methods developed for several decades, evaluation becomes crucial. The problem of defining a good segmentation remains unsolved and the solution mainly depends on the goal. A good segmentation can be defined as a segmentation true to one given by a human being.

The criteria of quantitative evaluation can be split into two classes, depending whether we possess or not a ground-truth which constitutes a reference segmentation. This reference is directly accessible in the case of computer generated images, but in the case of real images it must generally be built "by hand" by an expert of the application domain: layouts achieved by doctors, geographers, etc. with the help of computer-assisted drawing tools.

If we want to compare segmentation methods in a objective way, it is simpler to use synthetic images, for which a ground-truth is perfectly known, namely the segmentation which was used for synthetizing the image. The drawback of such a method is that these images do not represent all the possible situations of a real use.

This paper is based on "New Criteria for Evaluating Image Segmentation Results," by S. Philipp-Foliguet and L. Guigues, which appeared in the Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2006, May, 2006, Toulouse, France.

Although the evaluation on real images is certainly more realistic, it poses other problems, the main one being that, in a context where the number of regions to extract is not known a priori, there is generally no unique solution to the division of an image into "relevant" regions. The "relevance" of a region is indeed a notion highly dependent on the application. For example, working with an aerial image, someone who wants to separate the cultivated areas from the woods only needs "global" extraction of the fields, while someone who wants to establish a precise land-use classification requires individual delineation of each field. Hence the notion of "segmentation goal" is very important [1], a segmentation results cannot be evaluated without what Correia and Pereira call an "aplication scenario".

In this paper we are interested in what we call the "general partitioning problem", which is the problem where the number of regions in a solution is unknown a priori, the solutions space being the whole partitions space of the image (maybe restricted to the partitions with connected components). In this context, a key remark is that two human segmentations of the same image tend to be consistent in the sense that they are mutual refinements of each other [2]: some regions of a segmentation constitute an over-segmentation of some regions of the other one and conversely. In other words, the main difference between two human segmentations of the same image lies in the level of detail (we shall see some illustrations in section IV). If one attributes the differences in the segmentation results to the difference in the "goals" of the persons who made the segmentations, then one concludes that - in the general partitioning context - the main difference between two "segmentation goals" lies in the level of detail looked for in the segmentation. We thus claim that the notion of expected level of detail should explicitly be taken into account by segmentation evaluation criteria.

Yu and Shi [3] recently proposed a classification of segmentation methods in two big categories: on one hand "discriminative" approaches which, completely in the tradition of the unsupervised classification methods, consider the segmentation as a problem of grouping pixels into compact and well separated classes and on the other hand "generative model" approaches, which considers the segmentation as the inverse problem of finding an

hypothetical model which generated the data. The problem of evaluating a segmentation without reference is extremely close to the problem of segmentation itself. As for segmentation methods, we shall see that the criteria of evaluation without reference can be classified into two large categories: criteria of the first category adopt a "discriminative" approach and lead to some kind of contrast measures and criteria of the second category adopt a "generative model" approach and lead to the introduction of some underlying piece-wise model of the image and of an energy measuring the "quality" of this model. We have exploited this second approach and, starting from the multi-scale energy formulation of the segmentation of [4], [5], we propose new evaluation criteria which take explicitly into account the level of detail expected by the user. These criteria allow to rank the various competing segmentations according to a scale parameter and to reject the segmentations which are irrelevant for any scale.

Section II begins with a literature review of segmentation evaluation methods, separating the case where a ground-truth is available from the case where no ground-truth is available. By leaning on a criticism of the existing criteria, the section III then develops the new criteria proposed for the evaluation without reference. Before conclusion, section IV presents experiments carried out with these new criteria and comparisons with the main existing ones.

II. State of the art

Zhang [6] splits the segmentation evaluation methods into three groups : analytical methods, which analyse the segmentation process itself, "discrepancy methods", which measure differences with a reference image, and "goodness methods", which use quality measures "established according to human intuition" and "covering different aspects of an ideal segmentation".

Many criteria of "discrepancy" have been proposed, which can be used when a ground truth is available, usually given by an expert of the application domain, who is supposed to exactly know what he is expecting, in terms of accuracy, level of detail, etc. Among these discrepancy criteria let us cite Vinet measure [7], the measure of Yasnoff et al. [8] which counts the number of mis-segmented pixels, the Baddeley distance [9], and the ultimate measurement accuracy of Zhang [10]. The measure of consistency between segmentations of Martin [11] can also be used as a discrepancy measure between a segmented image and a reference image.

The problem is that the variability of a manual drawing, achieved by various experts is far from being negligible. It has been studied by Chalana and Kim in the case of medical images [12]. A big work to create a database of ground-truth images of photographs has been performed by D. R. Martin [11]. 1020 colour images issued from the Corel database have been manually segmented by about thirty persons, supplying 11 595 segmentations available on the internet (cf. Figs. 7 and 4). We will use them to validate our approach in section IV.

When there is no ground truth, a "goodness method" must be used, which can employ absolute quantitative criteria or consistency criteria between the segmentation results. Our approach falls in this category, and by the end of this section we expose the main existing approaches.

Correia and Pereira proposed a taxonomy of semantic criteria according to video segmentation scenario [1]. Of course some of these criteria involve mouvement information. But some of them are also valid for still images, they concern content complexity (spatial uniformity, regularity of shape and contrast between objects) and segmentation quality (precision of contours). The aim of our paper is exactly to quantify these criteria.

Also note that we are interested in comparing segmentation results as a whole. That is to say, we do not consider that the image contains some particular region of interest. For example in remote sensing or aerial images, if the goal is to produce maps, all areas must be properly extracted, with the same accuracy. We thus limit the following state of the art to criteria with the same aim : comparing segmentations as a whole.

Recently many criteria have been proposed for video segmentation, notably because of the standards imposed by MPEG-4 and MPEG-7 [13], [14], [15]. The criteria developed in this framework try to integrate perceptual metrics, like in [13], [15]. Motion is taken into account in these criteria, with the implicit hypothesis that moving objects are the most relevant. This explains that many criteria aim at describing the segmentation of an object rather than the global segmentation.

In this context, we present now the criteria which are the most used for comparing segmentation of images. According to the classification of segmentation methods proposed in [3], these criteria can be classified into two large categories: "discriminative" or "contrast" criteria and "generative model"-based criteria . The former ones express the quality through inter-region and/or intra-region variability measures, whereas the latter ones express the quality through the energy of a model which can include data-fitting terms and regularisation terms.

We first put the global notations of the section.

Let us consider an image I defined on set of sites X representing the spatial coordinates of the pixels (line, column) and a mapping f from X to Z. For example f can be the intensity for grey scale images (in this case Z is a subset of N) or the colour in one of the colour spaces for colour images (in this case Z is a subset of N^3).

In the following, we will note R a segmentation result to be evaluated. It is defined as a partition of X into non empty regions denoted $R_i, i = 1, ..., N$ verifing $R_i \cap R_j = \emptyset$ and $\bigcup_{i=1}^{N} R_i = X$. $A = |X|$ is the number of pixels of image I, and $A_i = |R_i|$ the number of pixels of region R_i. They are linked by the relation: $A = \sum_{i=1}^{N} A_i$.

We will note m_i (resp. σ_i) the mean (resp. the standard deviation) of f in region R_i.

A. Contrast criteria

1) Levine and Nazif inter-region contrast [16]: Let $c_{ij} = \frac{|m_i - m_j|}{m_i + m_j}$ be the contrast between two adjacent regions R_i and R_j.

The contrast of region R_i is then

$$c_i = \sum_{W_i} p_{ij} c_{ij}$$

where W_i denotes the indices of the regions adjacent to region i and $p_{ij} = \frac{l_{ij}}{l_i}$ is the ratio of the length of the common boundary between R_i and R_j to the perimeter of R_i.

The global contrast is then:

$$\frac{\sum_{R_i} w_i c_i}{\sum_{R_i} w_i}$$

w_i is a weight associated to each region, which can be the area of the region.

2) Zeboudj contrast [17]: This criterion takes into account both the intra-region contrast and the inter-region contrast.

Let $c(s,t) = \frac{|f(s) - f(t)|}{L-1}$ be the contrast between two pixels s and t, with f representing the intensity and L the maximum intensity.

The intra-region contrast for region R_i is:

$$I_i = \frac{1}{A_i} \sum_{s \in R_i} \max\{c(s,t), t \in W(s) \cap R_i\}$$

where $W(s)$ is a neighborhood of pixel s.

The extra-region contrast for region R_i is:

$$E_i = \frac{1}{l_i} \sum_{s \in F_i} \max\{c(s,t), t \in W(s), t \notin R_i\}$$

where F_i is the boundary of R_i and l_i the length of F_i.

The contrast of R_i is:

$$C(R_i) = \begin{cases} 1 - \frac{I_i}{E_i} & \text{if} \quad 0 < I_i < E_i \\ E_i & \text{if} \quad I_i = 0 \\ 0 & \text{else} \end{cases} \qquad (1)$$

The global contrast is then: $\frac{1}{A} \sum_i A_i . C(R_i)$.

This criterion was used in [17] to compare segmentations in regions on real and synthetic images. This criterion is not adapted to images too noisy or textured.

3) Rosenberger criterion [18]: To solve the problem of monochromatic images containing textures, Rosenberger proposes to characterize each region as textured or uniform region, thanks to a computation of intensity uniformity based on co-occurrence matrices.

Then he introduces intra-region disparity, denoted $\underline{D}$ and inter-region disparity, denoted $\overline{D}$. The former corresponds to the standard deviation of intensities for an uniform region and to a set of texture features for a textured region. $\overline{D}$ is defined as the difference between the means if both regions are uniform, and as the Euclidean distance between texture features if both regions are textured and to 1 if only one of the two regions is textured.

The global intra-region disparity is then defined as the weighted mean of the disparities computed for each region:

$$\underline{D} = \frac{1}{N} \sum_i \frac{A_i}{A} \underline{D}_i$$

and similarly for the global inter-region disparity.

Finally Rosenberger criterion is:

$$\frac{\overline{D} - \underline{D}}{2}$$

In Laurent et al. [19] most of these contrast-based criteria have been compared on a base of 100 synthetic images with or without texture. Zeboudj criterion and inter-region contrast of Levine and Nazif turned out to be the most efficient for all types of images except those including only textured areas, for which Rosenberger criterion was the most efficient. It is worth noting that, in spite of normalisation, some criteria have a very weak amplitude and also that some criteria give a low score to the ideal segmentation.

4) Contrast of Erdem et al. [14]: This criterion compares the mean colour of windows situated on either side of the boundary. Points are regularly put on the boundary and lines of lenght 3 or 5 are drawn from these points normal to the boundary on each direction. μ_i (resp. μ_o) is the mean colour vector of a window centered at the extremity of this line inside (resp. outside) the region, the colour contrast is mesured by :

$1 - \frac{1}{K} \sum_{i=1}^{K} \frac{\|\mu_i - . \mu_o\|}{\sqrt{3 \times 255^2}}$

where K is the number of points along the boundary and the colour vector is taken in Y, C_b ,C_r colour space.

As mentionned by the authors, when the boundary is properly estimated, this criterion is small, but on the contrary a small value does not necessarly imply a good localisation of the boundary.

B. Model-fitting criteria

1) Intra-region uniformity criterion of Levine and Nazif [16]: This simple criterion is based on the sums of the region variances. Being designed for texture-free images, it must thus be small.

$$\sum_i \sum_{s \in R_i} \left[f(s) - \frac{1}{A_i} \sum_{s \in R_i} f(s) \right]^2 = \sum_i \frac{\sigma_i^2}{C} \qquad (2)$$

f can be the intensity of pixel s or any other feature (colour, texture).

C is a normalisation coefficient, equal to the maximum possible variance:

$C = \frac{(f_{\max} - f_{\min})^2}{2}$

One can also weight each region by its area.

The advantage of this criterion is to be easily updated in case of region merging or splitting.

2) Dissimilarity measure of Liu and Yang [20]: This criterion is based on the number of regions, the area of the regions and their mean colour, in RGB space:

$$\frac{1}{1000 \times A}\sqrt{N}\sum_{i=1}^{N}\frac{e_i^2}{\sqrt{A_i}} \quad (3)$$

where e_i is the sum of the Euclidiean distances between the colour vector of pixels of region R_i and the colour vector assigned to region R_i in the segmented image (generally the mean colour of the region).

This criterion has to be small. Terms $\sqrt{N}$ at numerator (N is the region number)and $\sqrt{A_i}$ at denominator penalize the over-segmentation. It is close to Levine and Nazif criterion, the computation of the distances to the mean is slightly different, as well as normalisation, but the general idea is the same.

This dissimilarity measure was used by its authors [20] in order to find the best colour space for segmentation. Unfortunately, no space was found to be the best for all types of images. It will be used in our comparative tests of sections IV-B and IV-C .

3) Borsotti criterion [21]: The dissimilarity measure of Liu and Yang penalizes the segmentations with too many regions or with regions which are not homogeneous in terms of colour.

Borsotti $and\ al.$ proposed to improve it by this criterion:

$$\frac{1}{10000 \times A}\sqrt{N}\sum_{i=1}^{N}\left(\frac{e_i^2}{1+\log A_i}+\frac{N(A_i)^2}{A_i^2}\right) \quad (4)$$

where $N(A_i)$ is the number of regions with area equal to A_i.

This criterion must also be small. The first term of the sum favors homogeneous regions, as in Liu and Yang criterion. The second term has a high value when there are many small regions, which penalizes images segmented in many regions of the same size, especially if they are small.

It will be compared to other criteria in sections IV-B and IV-C.

Very few works are about the comparison of various evaluation criteria without reference. Chabrier et al. [22] overcome the problem by comparing the criteria with the Vinet measure, considered as a reference.

III. Multi-scale criteria of evaluation

A. The duality between segmentation and evaluation tasks

In this article, we focus on the case where there is no ground-truth and the goal is to estimate the relative quality of various segmentations of the same image.

This problem of evaluation without reference is extremely close to the problem of segmentation itself. Indeed, in a general way, the image segmentation problem amounts to formulate a quality function on the pairs $(image, partition)$ so that the expected partitions (the "acceptable" results of segmentation) obtain the best quality. This is supported by a principle of "comparison" which was proposed by Koepfler, Lopez and Morel [23]: "We shall adopt has principle without which no discussion about segmentation can even start, and which we call comparison principle. It states that given two different segmentations of a datum, we are always able to decide which of them is considered as better (or equivalent to) the other. Thus we assume the existence of some total ordering over all segmentations, and this can be simply achieved only if this ordering is reflected by some real functional E such that if $E(K_1) < E(K_2)$, then the segmentation K_1 has to be considered "better" than the segmentation K_2". In other words, segmentation and evaluation without reference both involve the definition of a "quality" measure Q on the pairs $(image, partition)$. For a segmentation task, given an image I, one looks for the partition P which maximizes $Q(I, P)$ over all the possible partitions of I. For an evaluation task, given an image I and a certain number of propositions of segmentation $P_1 \dots P_k$ of I, one looks for i which maximizes $Q(I, P_i)$, and claims that segmentation P_i is "the best". Hence, from a theoretical point of view, both problems of segmentation and of evaluation of segmentation without reference are identical and boil down to the definition of a suitable measure of quality Q. However, from a practical point of view, to obtain an effective algorithm of segmentation it is necessary to optimize the quality criterion (the energy) hence the choice of this criterion is mostly guided by the capacity to optimize it in reasonable time (exactly or approximately). Finally, what differentiates in practice these two dual tasks is that for evaluation purposes one can formulate more complex energies than for segmentation purposes because these energies only have to be calculated on a small number of segmentations and need not be optimized over the whole partition space.

B. Analysis of existing evaluation criteria

Indeed, when provided with an evaluation criterion, one can always view it as a segmentation criterion and evaluate the ability of the criterion to give acceptable segmentation results on some test images, simply asking: what solution(s) would we find if we optimized the criterion?

Also, in energy-based formulations of the segmentation problem, one systematically investigates whether the optimization problem is well-defined, in the Hadamard sense (does it have a solution? if so, is the solution unique? is it continuous with respect to the data?).

We thus propose to examine the evaluation criteria described in section **??** from this point of view.

Let us consider the case of an uniform image (whose pixels have the same value) and let us wonder which segmentation of this image would be considered as the best for each criterion. One easily verifies that all the criteria except that of Borsotti systematically return zero, whatever segmentation is proposed for the uniform image! The criterion of Borsotti awards the best note to the segmentation into a single region, what corresponds to

the expected result. In other words, all the criteria except that of Borsotti are ill-posed: they cannot decide in the simplest case of a uniform image.

General discussions on the ill-posedness of contrast-based (or discriminative) formulations of the segmentation problem can be found in [24], [25], [26], [4]. Ill-posedness of model-based (or generative model) formulations of the segmentation problem have been extensively discussed in the energy minimization-based segmentation community, see e.g. [27]. In next section, we review the general ideas which will lead us to the formulation of our multiscale evaluation criteria.

C. General energetical formulation and scale

1) Multiscale image segmentation: In a general way, the image segmentation problem can be thought of as a piecewise modeling problem (piecewise constant, polynomial, Gaussian, smooth, etc.) in which each region corresponds to a "piece" of the model. Once a class of model selected (for example piecewise constant), then - from the comparison principle - the search for the best model can always be formulated as an optimization problem: find a partition R of the image into regions and on each region R_i of R find the model M_{R_i} which minimizes a given total energy $E(R)$. Indeed, the energy has to take into account the fitness of the model by incorporating a term $E_D(R)$ which quantifies the distance between the model and the image. However, if we content ourselves with a goodness-of-fit energy, the optimal segmentation is eventually the absolute over-segmentation or something close. For example, if we consider piecewise constant models, then the model with one region per pixel, of value the one of the pixel, is always an exact solution to the problem, having a null distance to the image. Moreover, the problem is ill-posed: for the uniform image case, all partitions have a null distance.

To obtain a well-posed problem (and useful results), it is then necessary to incorporate an energy term $E_C(R)$, often called a "regularizer", and which penalizes too fine segmentations, that is too "complex" models. Considering independent models between regions, one then ends with energies which take the general form [5]:

$$E(k, R) = \sum_{R_i \in R} E_D(R_i) + k \times E_C(R_i) \qquad (5)$$

Where k is a real parameter which adjusts the relative contribution of the two energetical terms. Let us note that besides controlling the fineness of the solution, the "complexity" energy E_C also allows to control the geometrical regularity of the solution, for example by penalizing regions which have too ragged boundaries. In a probabilistic framework, the energy E_D can be viewed as the opposite of a log-likelihood of the data knowing the model and the energy $k \cdot E_C$ as the opposite of a log-prior probability of the model.

Let us remark that the Borsotti criterion, which was found above to be the unique well-posed criterion of the six criteria of section II, is a two terms-based energy of this kind.

In this framework, the choice of a segmentation depends on a compromise between goodness-of-fit and complexity of a model and there is intrinsically no best solution: certain applications can require a precise model - which thus will be complex - while other applications can require a rough - hence simple - description of the image.

If the energy E_C is an increasing function with respect to the fineness of the partitions, it is then shown in [4] that the parameter k of equation 5 allows to control the fineness of the solution, that is behaves as a *scale parameter*. As we already noticed, if $k = 0$ one gets a very tessellated model which perfectly fits the image; on the contrary, for sufficiently large values of k, the image is modeled by a single region.

2) Multiscale evaluation criteria: Based on the work on multiscale segmentation described in the previous section, we propose to address the dual problem of segmentation evaluation without reference by introducing a multiscale criterion of evaluation which is a *function* of the form:

$$k \longmapsto E_k(R) = \frac{1}{c} \sum_{i=1}^{N} E_D(R_i) + k \times \frac{1}{d} \sum_{i=1}^{N} E_C(R_i) \quad (6)$$

where E_D is a goodness-of-fit energy, E_C is a complexity energy, and the coefficients c and d are normalization coefficients which will be discussed below. k is the scale parameter, and $E_k(R)$ is an *affine* function of k.

As E_C is positive, $E_k(R)$ is an increasing function of k. Hence, consider two segmentations R and R' to compare. Two cases occur:

- Either $E_k(R) < E_k(R')$ for all k (or conversely), then the segmentation R is better than R' for all scales (or conversely),
- Or it exists k_0 such that $E_{k_0}(R) = E_{k_0}(R')$, then one of the two segmentations is better for small scales ($k < k_0$) and the other is better for large scale ($k > k_0$).

Hence, with this approach, segmentation results are not qualified in an absolute way. Assume one has two segmentations of the same scene: a coarse one and a fine one. Our multiscale criterion would elect the fine one for a task which requires a precise segmentation and the coarse one for a task which only requires a rough, global segmentation.

This analysis generalizes to an arbitrary number of segmentations $R_1 \dots R_n$. One easily shows that if E_C is positive then the set of scales for which a segmentation R_i is better than all other is an *interval*, possibly empty, which represents the range of scales for which this segmentation is the most relevant. Our approach thus allows to:

- Order the segmentations according to scale
- Eliminate the segmentations which are not relevant for any scale.

D. Various energy forms

Various forms of goodness-of-fit and complexity energy have been studied in [4]. We briefly review them here and describe the energies we used in our experiments.

1) Goodness-of-fit energies: Let R_i be a region of A_i pixels with values $X = (X_1, X_2, ..., X_{A_i})$. Let X_p^j be the j-st color component of pixel p. Let μ be the mean value of X and μ^j its j-st component. Let V be the covariance matrix of X, with general term:

$$V(j,k) = \frac{1}{A_i} \sum_{p=1}^{A_i} \left(X_p^j - \mu^j\right)\left(X_p^k - \mu^k\right)$$

The most basic goodness-of-fit energy is the one based on an underlying piecewise constant model and which quantifies the fit to the image using the L_2 norm:

$$Q(R_i) = \sum_{p=1}^{A_i} \|X_p - \mu\|^2 = A_i \cdot Tr(V)$$

where $Tr(V)$ is the trace of V. This energy has been first proposed by Mumford and Shah [27].

Another approach considers a piecewise Gaussian model. The values in a region are modeled as i.i.d samples of a Gaussian law. Generalizing to arbitrary dimensions the optimal encoding approach of Leclerc [28], [4] then ends up with an energy proportional to:

$$G(R_i) = \ln(\det V) = \sum_{j=1}^{3} \ln(\lambda_j)$$

where the λ_j are the eigenvalues of V.

Note that this definition only holds when the determinant is nonzero, which is not the case in many situations: regions with less than 3 pixels, uniform regions, regions with one uniform component... This can be solved by computing the determinant through the computation of the eigenvalues of V and thresholding the eigenvalues which are smaller than 1, for a color range in $[0, 256[$, see [4].

The computation of the covariance matrix can be done in any color space. In our experiments we simply used the RGB color space. Note however that both energies Q and G are invariant by rotation of the color space and that rescaling globally the image values does not change the order between segmentations but only produces a global homothety of the scale axis.

Other energy forms very close to the trace and the determinant can also be used:

$$Q'(R_i) = A_i \cdot Tr(\sqrt{V}) = A_i \sum_{j=1}^{3} \sqrt{\lambda_j}$$

$$D(R_i) = A_i \cdot \det V = A_i \prod_{j=1}^{3} \lambda_j$$

$$D'(R_i) = A_i \cdot \det \sqrt{V} = A_i \prod_{j=1}^{3} \sqrt{\lambda_j}$$

In the experiments presented below, we tested three of these energies: Q, G and D.

For images whose values are encoded in $[0, 2n[$ for each component, n^2 is an upper bound for the values of the covariances. Hence, the normalization coefficient c of equation 6 has been set to:

$$\begin{array}{ll} c = n^2 \times 3 \times A_i/100 & \text{for energy } Q \\ c = n^6 \times 3 \times A_i/10000 & \text{for energy } D \\ c = A_i & \text{for energy } G \end{array} \quad (7)$$

2) Complexity energy: The simplest form of complexity energy gives a constant energy to each region. Summing up on a partition, this leads to a global energy proportional to the number of regions of the segmentation.

Another possibility is to quantify the complexity of a region by the length of its boundary:

$$L(R_i) = \sum_{s \in \delta R_i} 1$$

where δR_i represents the boundary of region R_i. The total energy of a partition is then proportional to the total length of the interfaces in the partition.

One can also take into account the gradient magnitude along the boundaries, using:

$$LG(R_i) = \sum_{s \in \delta R_i} h(g(s))$$

where g represents the gradient magnitude and h is a positive decreasing function. The simplest choice is $h(x) = \frac{1}{\|x\|}$, which is valid as soon as the gradient magnitude is nonzero, which is normally the case on boundaries.

Other energies, based on polygonal approximations of the boundaries have been proposed in [4]: number of segments in the polygon, concavity, coherence of the orientations of the segments.

In the experiments presented below, we have only considered the energy proportional to the length of the boundaries. When the segmented image had contour pixels between regions, the computation of the energy was made by summing up these contour pixels. When the segmented image did not have contour pixels, the computation of the energy was made by summing up all the horizontal and vertical edges between adjacent regions.

The normalization coefficient d was set to the number of pixels in the image. The complexity term thus lies between 0 and 1.

IV. Experimental results

Our purpose is to provide users with a tool which allows to choose amongst various segmentation results according to a goal. Because he is an expert of his domain (medicine, geography, biology, etc.) the user knows the result he is expecting; in particular, he knows what level of detail he is expecting. And he is the only one able to evaluate if an algorithm result fits the expected segmentation or a manual segmentation.

	a	{d	b	c}	{e	f}
E_c	0.034	0.038	0.040	0.039	0.059	0.056
Q	17.41	16.03	15.06	15.14	7.59	9.67
D	0.0033	0.0058	0.0056	0.0050	0.0028	0.0043
G	15.14	14.69	14.68	14.61	13.36	13.96

TABLE I.
TABLE OF ENERGIES FOR EACH SEGMENTATION OF FIG. 1 (RANKED FROM COARSE TO FINE, ACCORDING TO HUMAN EXPERTS)

A. Comparison of energies

The proposed criterion is a weighted sum of two terms (cf. Eq. 6). We first computed separately both energy terms (goodness-of-fit and complexity) and we compared the various forms of goodness-of-fit energy. The aim is to check if the proposed criterion conforms to our visual perception.

We used the Berkeley dataset which contains 1000 images manually segmented by 30 human subjects, providing 12,000 hand-labeled segmentations [29]. As they have been drawn by human subjects, all theses segmentations can be considered as correct segmentations (at least for the author of the segmentation !). But the segmentations manually drawn by different persons sometimes present very various aspects (see Fig. 4) and these variations are often relative to the level of detail : some drawings only show the main regions, whereas some drawings go deeper in details. We have chosen 4 images of the Berkeley dataset in order to illustrate our method. They have been chosen because they are given with several segmentations of various levels of detail. We have asked several persons to compare the level of detail of each segmentation. The protocol followed recommandations ITU P.910 [30], subjects were asked to decide between two segmentations of the same image which one was the finest one, and which one the coarsest one or if it was impossible to decide. We could then obtain a partial ranking of the segmentations.

For example image #1 (Fig. 1) has been manually segmented by 6 human subjects. They ranked the various segmentations into two groups : (e) and (f) show accurate segmentations whereas the other ones are corser segmentations, (a) being the coarsest one.

We give in Table I the values of the complexity energy and of three forms of goodness-of-fit energies. We focused on Q, D, and G since the other ones (Q' and D') are very close to Q and D.

We notice that the values for the complexity energy are in conformity with the visual ranking : the coarser the segmentation, the smaller the complexity energy. For goodness-of-fit energy Q, the two clusters $\{$(e), (f)$\}$ on one hand and the others on the other hand are clearly extracted, they are less obvious with energy G and mismatched with energy D.

If we represent the couples (E_D, E_C) obtained for the six segmentations (cf. Fig 2), we can observe two or three clusters according to the goodness-of-fit energy. (e) and (f) form one cluster and the other results form one or two clusters, according to the energy. Only energy D clearly separates (a).

In Fig. 3a, the function $k \rightarrow E(k, R)$ is drawn for each segmentation result of Fig. 1 and for the goodness-of-fit energy Q. For small values of k (corresponding to a high level of detail), the straight line corresponding to segmentation (f) is under all others, meaning that if a fine segmentation is searched, (f) must be chosen. On the contrary for a coarse segmentation, (a) has the smallest total energy.

The same straight lines $k \rightarrow E(k, R)$ drawn with energy D indicates that (a) has the smallest energy, for all levels of details, which in not consistent with human ranking. And the goodness-of-fit energy G gives the same results as Q.

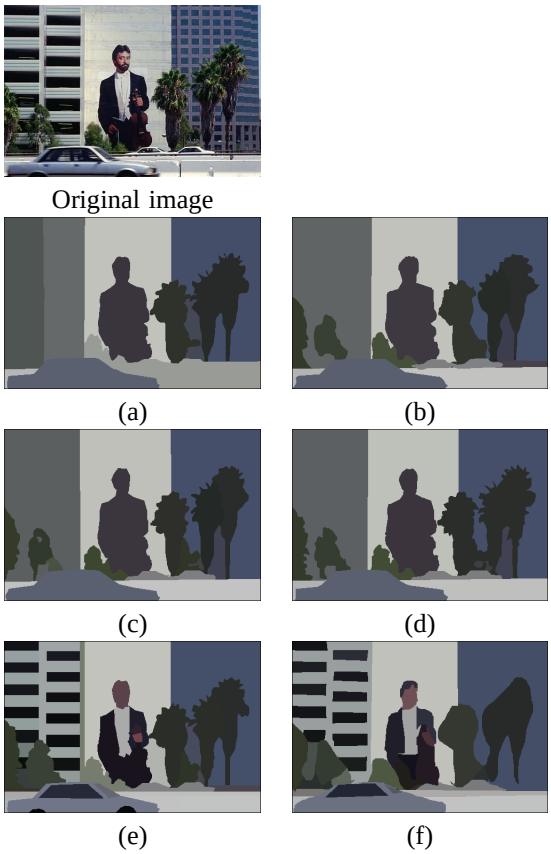
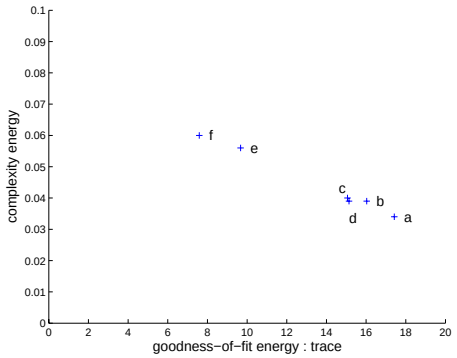


Figure 1. 6 manual segmentations of image #1 from Berkeley database

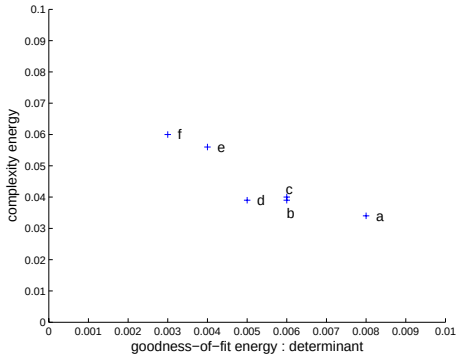
The second example is image #95 (Fig. 4) which is provided with 5 human segmentations. The result of the ranking was consensual: (a) and (b) on the one hand and (c) and (e) on the other hand have the same level of detail. (a) and (b) are visually very close, the drawing for image (a) seems more accurate for the frontier of the leaves on the head and on the legs, but on the contrary, the boat is not drawn. Human experts judged them as having the same level of detail. Finally they defined three groups, which are from the finest to the coarsest : (d), cluster $\{$(a), (b)$\}$, cluster $\{$(c), (e)$\}$.

We give in Table II the values of the complexity energy and of three forms of goodness-of-fit energies.

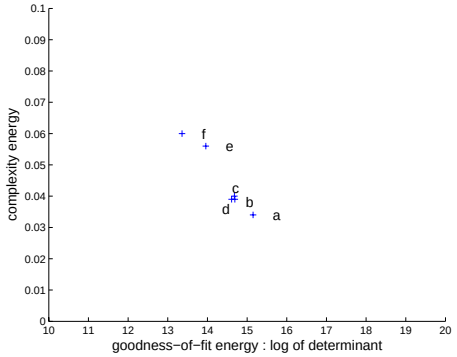
The values for the complexity energy are in conformity



(a) Goodness-of-fit energy : Q

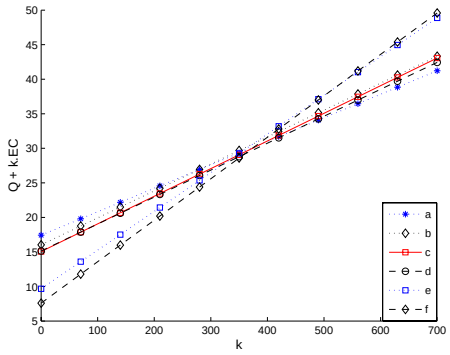


(b) Goodness-of-fit energy : D

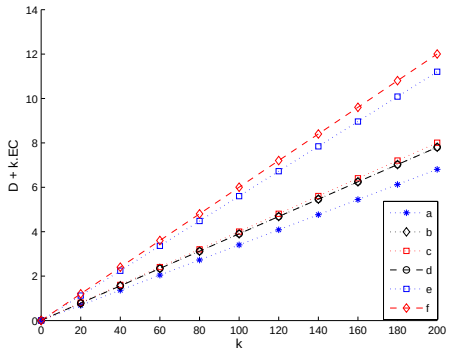


(c) Goodness-of-fit energy : G

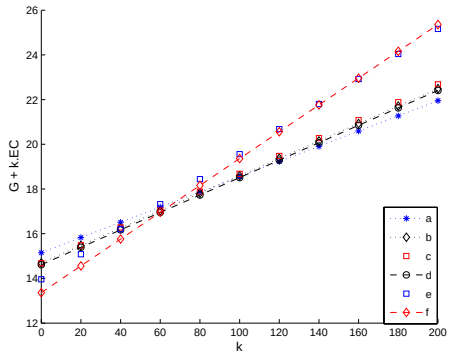
Figure 2. Edge energy versus internal energy for 5 segmentation results of image #1.



(a) Goodness-of-fit energy : Q



(b) Goodness-of-fit energy : D



(c) Goodness-of-fit energy : G

Figure 3. $k \rightarrow E(k, R)$ with three goodness-of-fit energies for the 5 manual segmentations of image #1

	d	a	b	c	e
E_c	0.051	0.042	0.042	0.039	0.040
Q	3.36	4.84	4.84	5.51	5.79
D	0.0033	0.0040	0.0051	0.168	0.306
G	11.05	12.70	12.52	12.90	12.39

TABLE II.
ENERGIES FOR EACH SEGMENTATION OF FIG. 4 (RANKED FROM FINE TO COARSE, ACCORDING TO HUMAN EXPERTS)

with the visual ranking : the finer the segmentation, the larger the complexity energy. Segmentations (a) and (b) which were judged as of the same level of detail have exactly the same complexity energy and the same value of energy Q, while their goodness-of-fit energies D and G lightly differ. (d) is judged by human beings as the finest one. Its complexity energy is the largest of the 5 results and its goodness-of-fit energy is the smallest for the three formula. (c) and (e) have beed judged as the coarsest ones and only goodness-of-fit energies Q and D rank them as worse than other images.

The plotting of couples (E_D, E_C) for the five segmentations of Fig 5, shows that they are clustered into three clusters ((a) and (b) are surimposed) with goodness-of-fit

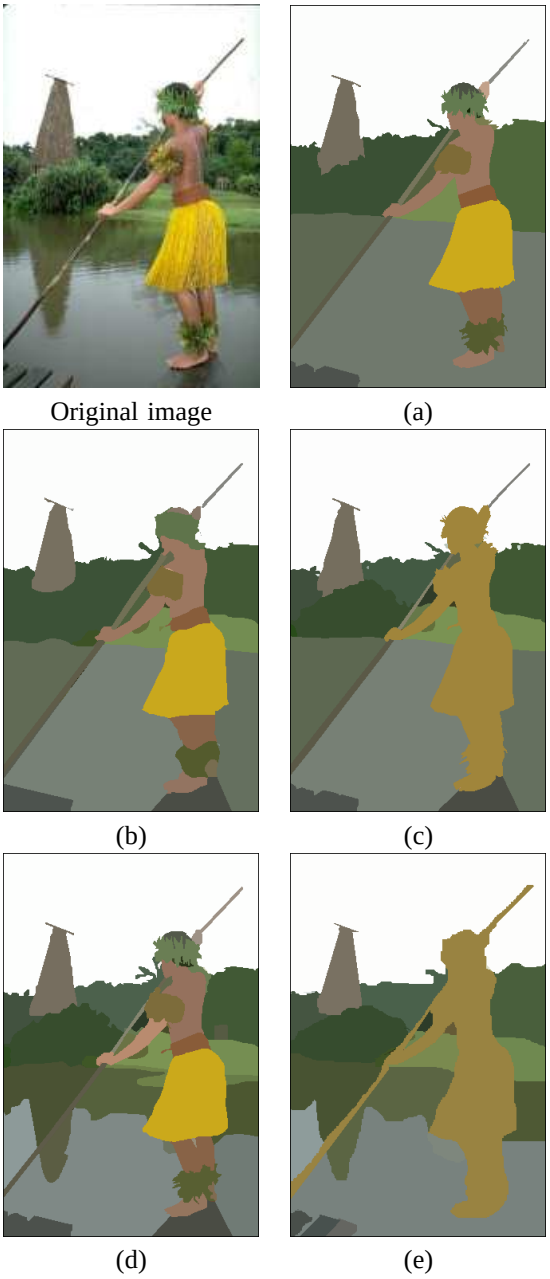
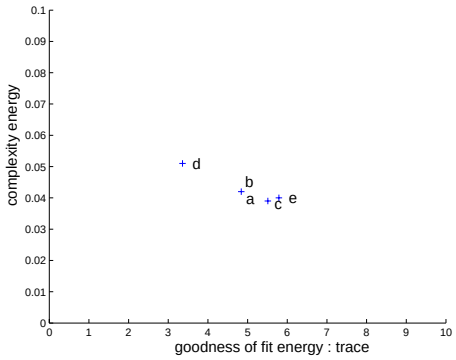
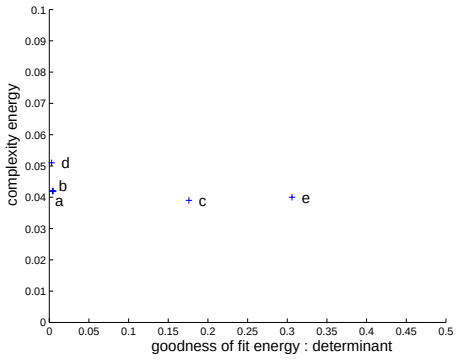


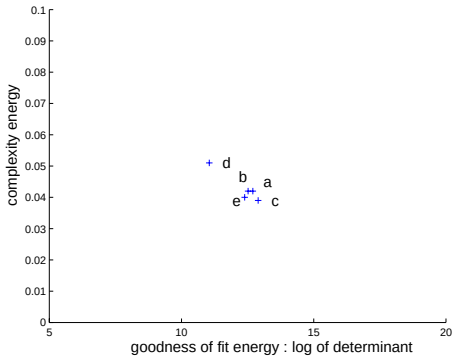
Figure 4. 5 manual segmentations of image #95 from Berkeley database



(a) Goodness-of-fit energy : Q



(b) Goodness-of-fit energy : D



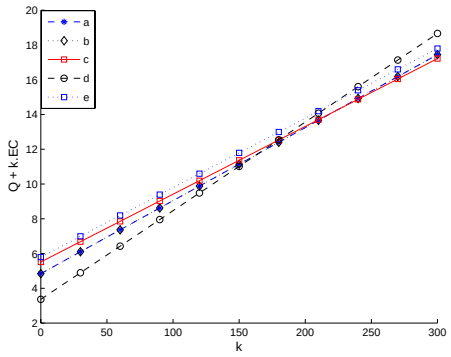
(c) Goodness-of-fit energy : G

Figure 5. Edge energy versus internal energy for 5 manual segmentation of image #95

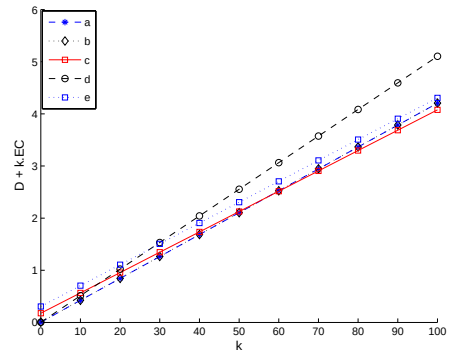
energy Q, wheras for the other energies, the discrimination between images is less obvious.

If we draw functions $k \rightarrow E(k, R)$ for each segmentation result of this image and for the goodness-of-fit energy Q (Fig. 6a), we notice that for all values of k, the dot lines representing segmentations (c) and (e) are very close to each other and almost parallel, (c) being always lower than (e). This can be interpreted as : although these two segmentations have almost the same level of detail, (c) more fits the edges of the regions (both complexity and goodness-of-fit energies are lower for (c) than for (e)). If a fine segmentation is searched, one has to prefer segmentation (d), since the straight line representing (d) is under any other lines for small values of k. For a coarse segmentation, (c) is the best one (smaller values for large values of k), and for an intermediate level, (a) or (b) are the best ones. The same straight lines $k \rightarrow E(k, R)$ drawn with energies D and G are more difficult to interpret, since values for D are very close to each other, and with energy G, it clearly gives result (d) as the finest one, but has difficulties to separate the other ones.

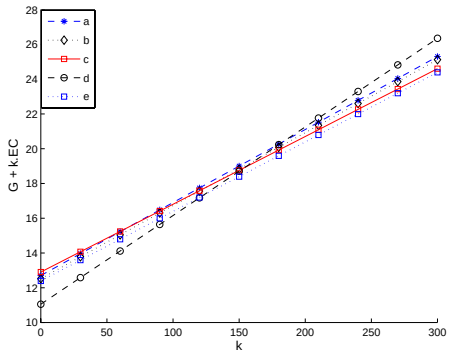
From all the tests we made, we can confirm that the goodness-of-fit Q is more discriminative and more conform to the visual ranking given by human subjects. We give other results in Fig. 8, corresponding to the manual segmentations of image #16 (Fig. 7). The human experts ranked the 5 results as follows, from the finest to the coarsest : (e), (d), cluster $\{$(b), (c)$\}$, and (a) and the



(a) Goodness-of-fit energy : Q

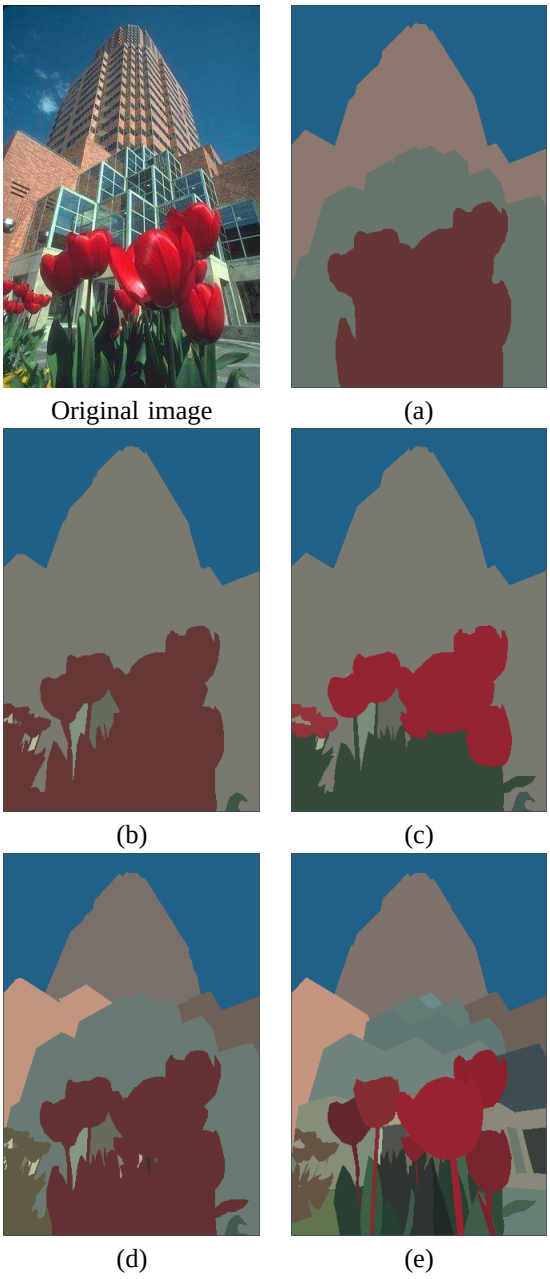


(b) Goodness-of-fit energy : D



(c) Goodness-of-fit energy : G

Figure 6. $k \rightarrow E(k, R)$ with three goodness-of-fit energies for the 5 manual segmentations of Image #95



Original image (a) (b) (c) (d) (e)

Figure 7. 5 manual segmentations of image #16 of Berkeley database.

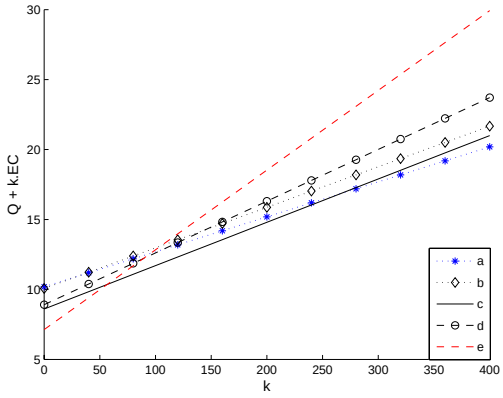
straight lines with goodness-of-fit energy Q indicates to choose (e) for a fine segmentation,(c) for an intermediate segmentation and (a) for a coarse segmentation. D gives (a) as the best at any level of detail and G gives the same results as Q.

Image #21 is provided with 7 manual segmentations (cf Fig.9), ranked by human experts as (from fine to coarse) : (a), (b), (c), (d) and cluster $\{$(e), (f), (g)$\}$. The energy criterion (with goodness-of-fit energy Q indicates that for a fine segmentation, (a) gives the best result and for a coarse segmentation, (g) or (f) are the best choices.

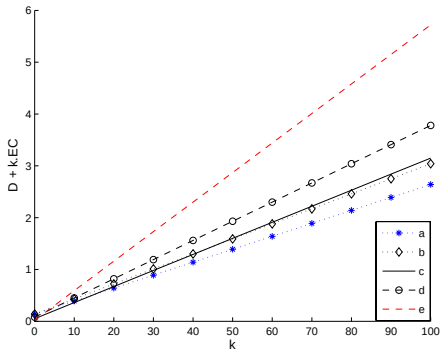
B. *Comparison of evaluation criteria*

We now compare our multiscale criterion, with goodness-of-fit energy Q (denoted by MS) to the main criteria used in the litterature : Levine and Nazif, Liu and Yang, Borsotti et al. on the images for which we have a ground truth. We report here the results for the same 4 images as in the previous section. As our criterion is multiscale, we give the results for two different scales $k = 10$ and $k = 1000$, which give the fine level of detail for $k = 10$ and a coarse level of detail for $k = 1000$.

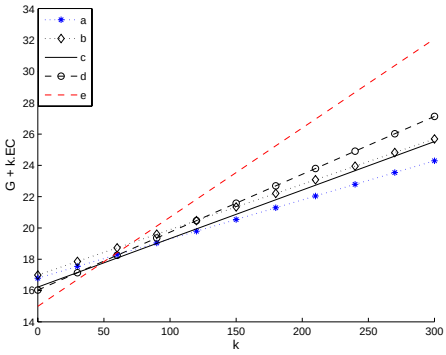
Criteria different from MS always choose the segmentation with the smallest number of regions as the best one, they more or less rank results according to the number of regions. As shown previously, our criterion allows to



(b) Goodness-of-fit energy : Q



(b) Goodness-of-fit energy : D



(c) Goodness-of-fit energy : G

Figure 8. $k \rightarrow E(k, R)$ for three goodness-of-fit energies for the 5 manual segmentations of Fig. 7.

chose a result according to a particular level of detail. For example for Image #95 (cf. Table IV) (d) is considered as the best one for a fine segmentation, and (c) for a coarse segmentation.

C. Comparison of evaluation criteria on algorithm results

We dispose of several segmentation results obtained from various algorithms. Image House is widely used in the image processing community to compare algorithms, it includes textured and non-textured parts. We used the 5 segmentation results published in [31] to illustrate colour image segmentation methods and a fuzzy method (F)

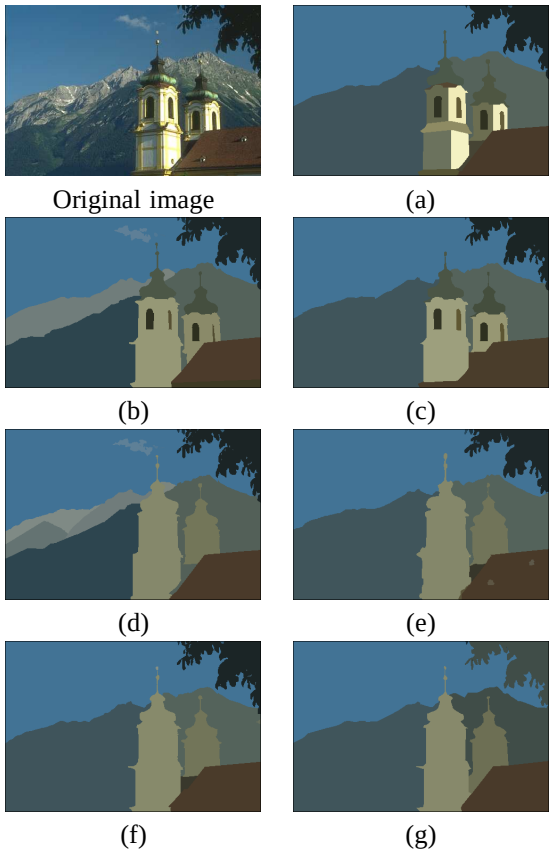


Figure 9. 7 manual segmentations of image #21 from Berkeley database

	a	b	c	d	e	f
# regions	14	30	23	19	58	63
Levine-Nazif	**5.39**	7.02	6.63	6.63	10.68	12.97
Liu -Yang	**0.27**	0.39	0.32	0.32	0.49	0.61
Borsotti	**0.31**	0.41	0.34	0.34	**0.31**	0.40
MS $k = 10$	17.4	16.4	15.4	15.4	**8.2**	10.2
MS $k = 1000$	**51.9**	54.6	54.6	54.6	67.4	65.3

TABLE III.
COMPARISON OF CRITERIA FOR THE 7 MANUAL SEGMENTATIONS OF IMAGE #1 (FIG. 1)

	a	b	c	d	e
# regions	33	69	48	72	43
Levine-Nazif	9.5	10.9	**6.25**	15.68	9.02
Liu -Yang	**0.16**	0.26	0.20	0.25	0.18
Borsotti	**0.14**	0.20	0.18	0.15	0.18
MS $k = 10$	5.26	5.26	5.90	**3.87**	6.19
MS $k = 1000$	46.7	47.0	**45.0**	54.8	45.7

TABLE IV.
COMPARISON OF CRITERIA FOR THE 5 MANUAL SEGMENTATIONS OF IMAGE #95 (FIG. 4)

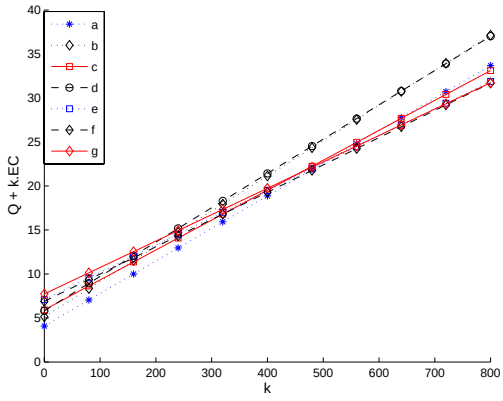


Figure 10. $k \rightarrow E(k, R)$ for goodness-of-fit energies Q for the 7 manual segmentations of Fig. 9.

	a	b	c	d	e
# regions	4	11	18	27	67
Levine-Nazif	**2.76**	4.56	6.31	8.78	14.9
Liu and Yang	**0.06**	0.10	0.11	0.20	0.46
Borsotti	**0.1**	0.17	0.17	0.26	0.32
MS $k = 10$	10.45	10.36	8.91	9.29	**7.70**
MS $k = 100$	12.7	13	**11.73**	12.65	12.8

TABLE V.
COMPARISON OF CRITERIA FOR THE 5 MANUAL SEGMENTATIONS OF IMAGE #16 (FIG. 7)

[32] (cf. Fig. 11). The 5 former results were respectively obtained by split and merge algorithm (SM), Tominaga (T), competitive learning (C), region growing (G), 2D histogram classification (H). Visually, Split and merge is not accurately segmented since blocks are visible. Fuzzy is not accurate as well since edges are not very straight, but the region number is closer to the number we visually perceive than for other results (see Table VII). T, C and H include many very small regions and they look very similar to each other. G also includes very small regions but less numerous than the three previous results, and mostly located on the edges.

At first analysis, segmentation G seems the most relevant, but a deeper analysis shows that there are many useless regions, and thus F is also a good solution. We compared the different energy forms of the proposed multiscale criterion on several results of segmentation, obtained by various algorithms.

	a	b	c	d	e	f	g
# regions	28	33	19	39	26	24	12
Levine- Nazif	5.42	3.95	3.93	3.92	4.12	3.03	**2.66**
Liu- Yang	0.13	0.16	0.12	0.17	0.15	0.14	**0.10**
Borsotti	**0.11**	0.14	0.12	0.17	0.17	0.16	0.13
MS $k = 10$	**4.45**	5.54	6.26	6.21	7.39	7.20	8.04
MS $k = 1000$	41.31	45.12	39.70	45.25	38.39	38.38	**37.74**

TABLE VI.
COMPARISON OF CRITERIA FOR THE 7 MANUAL SEGMENTATIONS OF IMAGE #1 (FIG. 9)

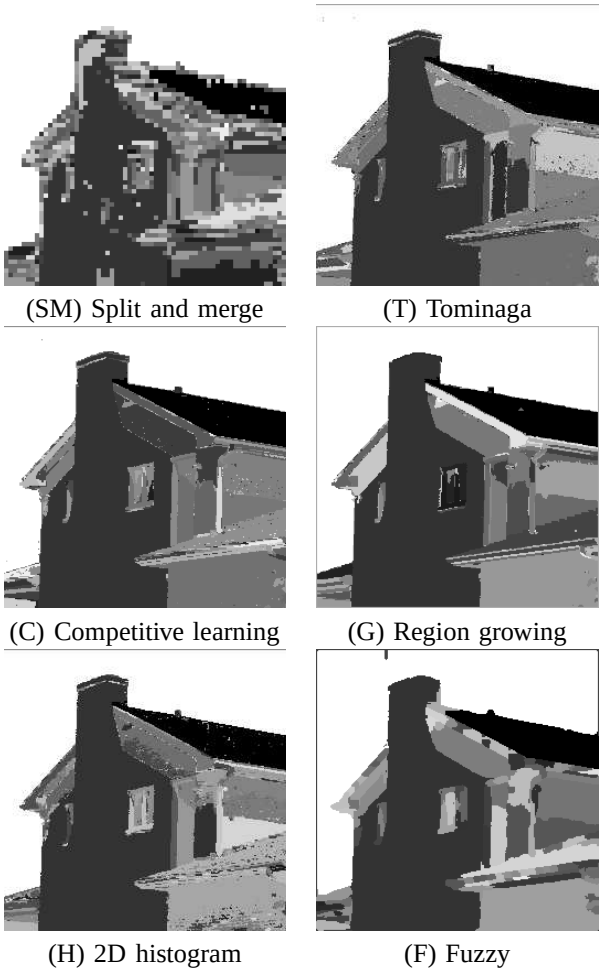


Figure 11. 6 segmentation results of image House

On the representation (E_D, E_C) (Fig. 12) with goodness-of-fit energy Q, one can observe that Split and merge on one hand and Fuzzy on the other hand are well discriminated from other results.

In Fig. 13, the function $k \rightarrow E(k, R) = Q(R) + k.E_C(R)$ is drawn for each segmentation result of Fig. 12. For all values of k, the straight line representing result G is always lower than those representing results C, T, SM and H. Hence G is always better than these four methods, whatever the scale (or level of detail) is. The comparison of F and G depends on the expected level of detail : for a coarse segmentation (large k), F is better than G, and conversely for a fine segmentation (small k). From this graphics, one can conclude that, among the 6 segmentation results we have, if we look for a coarse segmentation ($k > 25$) of image House, F gives the best segmentation, and for a finer resolution, G gives the best result.

All these criteria (except Borsotti) give close values for images visually close : T, C, H for the House obtain scores of the same order. The Borsotti criterion is extremely sensitive to the regions of one or two pixels, because of the second term of Eq. 4.

Another frequently used image is image Parrot, for

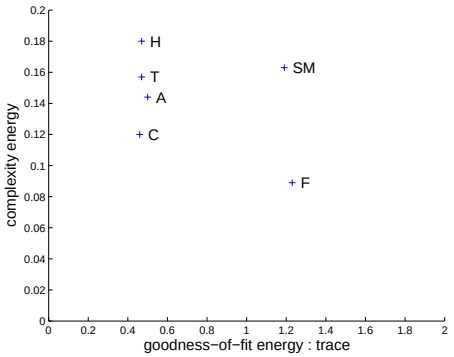


Figure 12. complexity energy versus goodness-of-fit energy : Q for 6 segmentation results of image House.

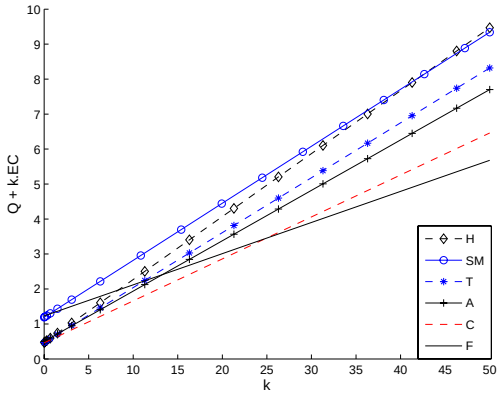


Figure 13. $k \rightarrow E(k, R)$ with internal energy Q for the 6 segmentations R of Fig. 11.

which we have 5 segmentation results (cf. Fig. 14). The first 4 were obtained by various classification algorithms : fuzzy C-means (FCM), a neuro-fuzzy method (NF), Kohonen method (KO) and k-means (KM) [33]. The last one was obtained by fuzzy segmentation (F) [32]. Visually, the classification methods over-segment, whereas the fuzzy method under-segment. Moreover the results of methods FCM, KO and NF seem very similar. This similarity is well conveyed on Fig. 15, with goodness-of-fit energy Q : we find the three beforehand mentioned groups, the fuzzy classifications on one hand, the fuzzy segmentation on the other hand and k-means in a third group.

	SM	T	C	G	H	F
# regions	379	968	667	379	994	97
Levine-Nazif	116	78	65	49	70	**31**
Liu-Yang	3.2	0.40	0.37	**0.25**	0.39	0.47
Borsotti	0.4	29	8	1.1	24	**0.1**
MS $k = 10$	2.82	2.04	1.94	**1.66**	2.28	2.12
MS $k = 100$	17.5	16.2	14.9	12.5	18.5	**10.2**

TABLE VII.
COMPARISON OF CRITERIA FOR THE 6 SEGMENTATION RESULTS OF IMAGE HOUSE (BOLD : THE BEST RESULT ACCORDING TO EACH EVALUATION CRITERION)

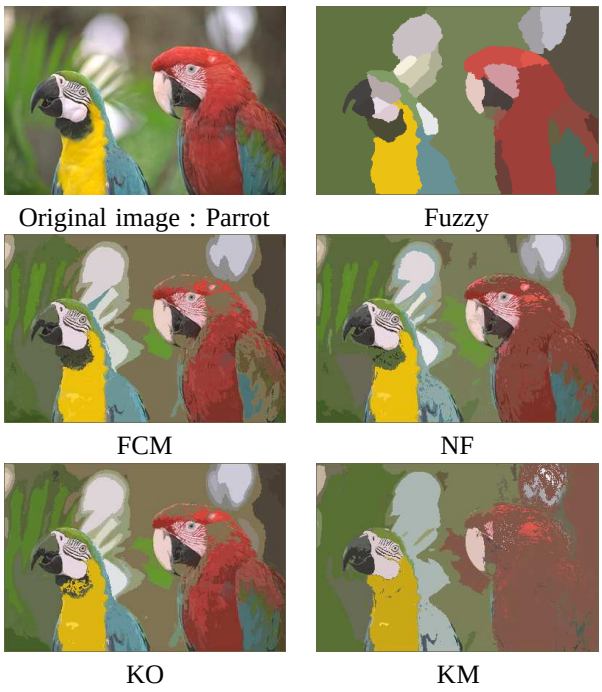


Figure 14. 5 segmentation results for image Parrot, each region labeled with its mean color

Fig. 15 clearly shows the similarity between the 3 results FCM, NF et KO whatever the scale. k-means is interesting at no scale. For a coarse scale ($k > 10$) the Fuzzy method gives the best result.

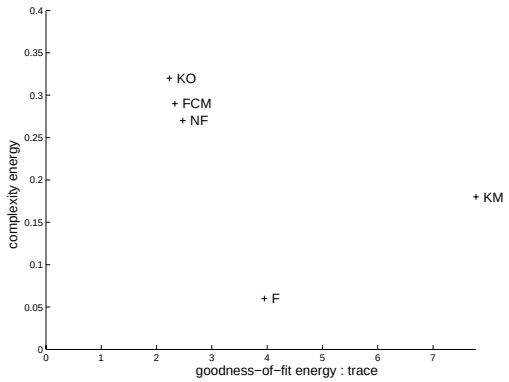


Figure 15. Complexity energy versus Goodness-of-fit energy : Q for 5 segmentation results of image Parrot.

As mentioned beforehand, parameter k in Eq. 5 sets the scale or the expected level of detail of the segmentation. The algorithms which provide few regions are favored by the complexity energy but disadvantaged by the goodness-of-fit energy.

The complexity energy is very linked to the region number, which explains why when k is large, the results with many regions have a small score.

It is hard to tune k in an abolute way. The number of regions depends on the size, on the content of the image and on the expected level of detail. The values of all the criteria only have a relative meaning. They have to be

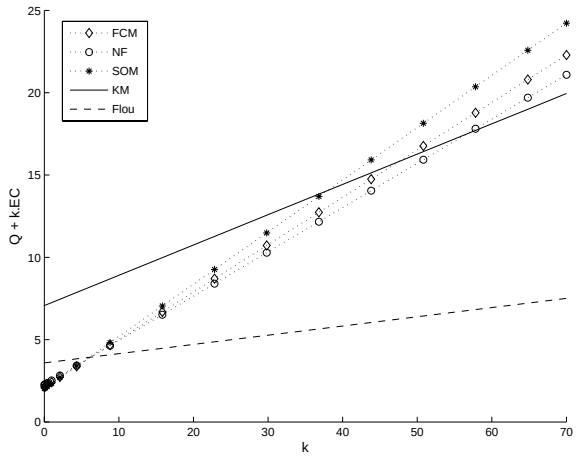


Figure 16. $k \rightarrow E(k, R)$ with goodness-of-fit energy Q for the 5 segmentations results of image Parrot.

	FCM	NF	KO	KM	F
# regions	2952	2748	3398	1708	34
Levine-Nazif	95	49	57	18	**7**
Liu-Yang	1.18	1.06	1.29	1.36	**0.18**
Borsotti	198	171	360	38	**0.11**
MS $k = 3$	3.0	3.06	**2.96**	7.51	3.74
MS $k = 100$	30.96	29.11	33.70	25.42	**9.21**

TABLE VIII.
COMPARISON OF CRITERIA FOR THE 5 SEGMENTATION RESULTS OF IMAGE PARROT (BOLD : THE BEST RESULT ACCORDING TO EACH EVALUATION CRITERION)

compared to each other, they have no unit. In the same way, scale k for our criterion cannot be tuned. The way for using the criterion is to draw the straight lines for each result. If a straight line is always below another one, the corresponding segmentation is not interesting (at no scale), the complexity of the edge is not compensated by the fitting of the region to the initial image. As a genaral rule, if an over-segmentation (or a fine segmentation) is preferred, k must be small, less than 10. On the contrary if a coarse segmentation, with few regions, is preferred, k must be chosen large, for example larger than 100.

V. CONCLUSION

Based on the remark of a duality between image segmentation tasks and segmentation evaluation tasks without reference, we have shown that the main existing evaluation criteria are ill-posed once cast into segmentation problems. Based on the previous work on multiscale image segmentation of [4], [5] we have related this ill-posedness to the fact that well-posed criteria must at least incorporate two antagonist terms, a goodness-of-fit term and a complexity term, and that the balance between the two terms rules the level of detail in the expected segmentation result. We have thus proposed to explicit this scale parameter (which was implicitly present but "hardly set" in the unique well-posed criterion among those examined, namely Borsotti criterion) and to let the user set this parameter according to his goal, thus ending with a multi-scale criterion. Indeed, it is illusory to try to compare a coarse segmentation which only delineates the global units of a scene to a fine segmentation which delineates every small detail. None of the two is *absolutely* better than the other. However, the former is better for coarse segmentation tasks whereas the latter is better for fine ones. Hence the need to explicit in evaluation criteria a parameter which models the expected level of detail in the segmentation result. Our experiments show that the proposed criteria finally allows to sort competing segmentation results with respect to their level of detail and to reject the segmentations which are irrelevant for all scales.

ACKNOWLEDGMENT

The authors wish to thank Ludovic Macaire and Fella Hachouf for their segmentation results respectively on the image House and the image Parrot, as well as David Picard and Jérôme Dantan for their evaluation softwares.

REFERENCES

[1] P. L. Correia and F. Pereira, "Classification of video segmentation application scenarios," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 14, no. 5, pp. 735–741, May 2004.

[2] J. S. Cardoso and L. Corto-Real, "A measure for mutual refinements of image segmentations," *IEEE trans. on Image Processing*, Accepted, to appear.

[3] S. Yu and J. Shi, "Segmentation given partial grouping constraints," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 2, pp. 173–183, 2004.

[4] L. Guigues, "Modèles multi-échelles pour la segmentation d'images," Ph.D. dissertation, Université de Cergy-Pontoise, 2003.

[5] L. Guigues, J.-P. Cocquerez, and H. Le Men, "Scale-sets image analysis," *International Journal of Computer Vision*, vol. 68, no. 3, pp. 289–317, July 2006.

[6] Y. J. Zhang, "A review of recent evaluation methods for image segmentation," in *Signal Processing and its Applications, Sixth International, Symposium on.*, vol. 1, 2001, pp. 148–151.

[7] L. Vinet, "Segmentation et Mise en Correspondance de Régions de Paires d'Images Stéréoscopiques," Ph.D. dissertation, Université Paris IX - Dauphine, July 1991.

[8] W. A. Yasnoff, W. Galbraith, , and J. Bacus, "Error measures for objective assessment of scene segmentation algorithms," *AQC*, vol. 1, pp. 107–121, 1979.

[9] D. L. Wilson, A. J. Baddeley, and R. A. Owens, "A new metric for grey-scale image comparison," *Int. J. of Computer Vision*, vol. 24, pp. 5–17, 1997.

[10] Y. J. Zhang, "Evaluation and comparison of different segmentation algorithms," *Pattern Recognition Letters*, vol. 18, no. 10, pp. 963–974, 1997.

[11] D. R. Martin, "An empirical approach to grouping and segmentation," Ph.D. dissertation, University of California, Berkeley, USA, 2002.

[12] V. Chalana and Y. Kim, "A methodology for evaluation of boundary detection algorithms on medical images," *IEEE Trans. on Medical Imaging*, vol. 16, no. 5, pp. 642–652, 1997.

[13] E. Drelie Gelasca and T. Ebrahimi, "On Evaluating Metrics For Video Segmentation Algorithms," in *VPQM 2006 Second International Workshop on Video Processing and Quality Metrics for Consumer Electronics*, ser. Parallel Computing in Electrical Engineering. http://www.intel.com: INTEL, 2006.
[14] C. Erdem Eroglu, B. Sankur, and M. Tekalp, "Performance measures for video object segmentation and tracking," *IEEE Image Processing*, vol. 13, no. 7, pp. 937–950, July 2004.
[15] P. L. Correia and F. Pereira, "Stand-alone objective segmentation quality evaluation," *EURASIP J. Appl. Signal Process.*, vol. 2002, no. 4, pp. 389–400, 2002.
[16] M. Levine and A. Nazif, "Dynamic measurement of computer generated image segmentations," *IEEE Trans. on PAMI*, vol. 7, no. 25, pp. 155–164, 1985.
[17] J.-P. Cocquerez and S. Philipp, *Analyse d'images: filtrage et segmentation*. Paris: Masson, 1995.
[18] S. Chabrier, C. Rosenberger, H. Laurent, B. Emile, and P. March, "Evaluating the segmentation result of a gray-level image," in *Proc. of 12th EUSIPCO, Vienne, Austria*, 2004.
[19] H. Laurent, S. Chabrier, C. Rosenberger, B. Emile, and P. Marché, "Etude comparative de critères d'évaluation de la segmentation," in *19th GRETSI*, Paris, France, june 2003.
[20] J. Liu and Y.-H. Yang, "Multiresolution color image segmentation," *IEEE Trans. on PAMI*, vol. 16, no. 7, pp. 689–700, 1994.
[21] M. Borsotti, P. Campadelli, and R. Schettini, "Quantitative evaluation of color image segmentation results," *Pattern Recognition Letters*, vol. 19, pp. 741–747, 1998.
[22] S. Chabrier, B. Emile, C. Rosenberger, and H. Laurent, "Unsupervied performance evaluation of image segmentation," *EURASIP Journal on Applied Signal Processing, Special Issue on Performance Evaluation in Image Processing*, 2006.
[23] G. Koepfler, Lopez, and J.-M. Morel, "A multiscale algorithm for image segmentation by variational method," *SIAM journal on numerical analysis*, vol. 31, no. 1, pp. 282–299, 1994.
[24] J. Shi and J. Malik, "Normalized Cuts and Image Segmentation," *IEEE trans. on PAMI*, vol. 22, no. 8, pp. 888–905, August 2000.
[25] P. Felzenszwalb and D. Huttenlocher, "Efficiently computing a good segmentation," 2001.
[26] L. Guigues, H. L. Men, and J.-P. Cocquerez, "The hierarchy of the cocoons of a graph and its application to image segmentation," *Pattern Recognition Letters*, vol. 24, no. 3, pp. 1024–1066, 2003.
[27] D. Mumford and J. Shah, "Optimal approximations by piecewise smooth functions and asociated variational problems," *Comm. Pure Appl. Math.*, vol. 42, pp. 577–685, 1989.
[28] Y. Leclerc, "Constructing simple stable descriptions for image partitioning," *Int. J. of Computer Vision*, vol. 3, no. 1, pp. 73–102, 1989.
[29] 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," in *Proc. 8th Int'l Conf. Computer Vision*, vol. 2, July 2001, pp. 416–423.
[30] ITU P.910, "Subjective video quality assessment methods for multimedia applications recommendation p.910," in *Int. Telecommunication Union*, Geneva, Switzerland, 1996, pp. 573–579.
[31] A. Trémeau, C. Fernandez-Maloigne, and P. Bonton, *Image numérique couleur*. Paris: Dunod, 2004.
[32] S. Philipp-Foliguet, M. B. Vieira, and M. Sanfourche, "Fuzzy segmentation of color images and indexing of fuzzy regions," in *First Europ. conf. on Colour in Graphics, Imaging and Vision*, Poitiers, France, 2002, pp. 507–512.
[33] M. B. N. Mezhoud, F. Hachouf, "Segmentation dimages couleurs par une méthode neuro-floue," in *Taima03*, Hammamet, Tunisie, 2003, pp. 170–176.

Sylvie Philipp-Foliguet is Professor at the National School of Electronics (ENSEA) of Cergy-Pontoise since 1988. She manages the MIDI (Multimedia Indexing and Data Integration) team of laboratory ETIS (CNRS).

Her research domains are image segmentation and interpretation. Concerning the first topic she works more particularly on the fuzzy approaches of the segmentation and the performance evaluation. The applications of the second topic concern image indexing and retrieval from databases ; she developed methods using fuzzy regions, graph matching and statistical learning.

Laurent Guigues received his PhD degree in computer science, signal and image analysis, from the University of Cergy-Pontoise in 2003.

He is currently working for the french national research agency (CNRS) at CREATIS-LRMN laboratory in Villeurbanne, France. His research interests include image segmentation and evaluation, medical image analysis and Monte Carlo simulation for nuclear imaging and cancer therapy.