NEW CRITERIA FOR EVALUATING IMAGE SEGMENTATION RESULTS

Sylvie Philipp-Foliguet
ETIS, CNRS UMR 8051
6 avenue du Ponceau, Cergy-Pontoise, France
philipp@ensea.fr

Laurent Guigues
CREATIS, CNRS UMR 5515, Inserm U 630
INSA, 7 rue Jean Capelle, Villeurbanne, France
laurent.guigues@creatis.insa-lyon.fr

ABSTRACT

This paper is dedicated to evaluation criteria for image segmentation when there is no ground truth. New criteria based on a formulation of the image segmentation as a piecewise modeling problem are proposed. These criteria take into account both the complexity of the segmented image, through the total boundary length and the goodness-of-fit through a distance between model and initial image. They allow to rank segmentation results or human segmentations, according to an expected level of detail. These new evaluation criteria are compared to the most used evaluation criteria both on results of segmentation algorithms and on manual segmentation achieved by humans.

1. 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. But on one hand, in many cases, no human segmentation is available, and on the other hand, the human segmentation can change from one user to another, accordingly for example to the application. The variability between people [1] often lays in the accuracy of the segmentation, which depends both on the position of the detected edges and on the number of regions, which is linked to the expected level of detail of the result.

In this paper, we focus on defining evaluation criteria for segmentation of colour images into regions when no ground truth is available, which is the most general and the most difficult problem.

Numerous criteria have been proposed, aiming at quantifying quality or legibility of the segmented image. When there is no ground truth, they consist in a single quality criterion, measuring the uniformity inside regions [2], or the adequacy to a model, including data-fitting and complexity [3] [4]. We propose to better model both the adequacy to data and the simplicity of the segmentation and to make explicit the balance between both terms, by the way of a parameter acting as a scale parameter.

2. PROPOSITION FOR NEW EVALUATION CRITERIA

In a general way, the segmentation can be thought of as a problem of piecewise modeling of the image : constant,or polynomial, or gaussian, ..., modeling. Once the type of model chosen (for example piecewise constant), the research of the best model can be formulated as an optimization problem : to find out both a partition P of the image into regions, and for each region R of P, the model M_R, which minimizes a total energy E. The energy takes into account the modeling quality, through a term $E_D(P)$, which measures the distance between the model and the image. With this term of adequacy to data, the optimal segmentation is obtained with the absolute over-segmentation or a similar partition. For example, if we consider a piecewise constant model, the exact solution is the segmentation with one region by pixel, with value the colour of the pixel. Of course this result is never useful, over-segmentation must be penalized. To do this, a term of "complexity" $E_C(P)$ is added, which aims at avoiding too fine segmentations, but also segmentations with too tortuous region boundaries. If we consider models independent for each region, we obtain an energy of general form [5] :

$$E(k, P) = \sum_{R \in P} E_D(R) + k \cdot E_C(R) \qquad (1)$$

where k is a real parameter, which tunes the relative contribution of the two energy terms. Within this framework, the choice of a segmentation results from a compromise between goodness-of-fit and complexity of the model. There is no intrinsic best solution : some applications need a coarse description of the image, while others need a precise model, which is thus more complex. Energy E_C is a function growing with the sharpness of the partition, parameter k controls the sharpness of the solution, that is to say it behaves as a *scale parameter* : if $k = 0$, the best model is a

very divided model, which perfectly fits to image, while for a large enough value of k, the image is modeled by a single region. Following this idea, it has been proposed in [5] not to choose one partition minimizing Eq. (1) for a value of k set up a priori, but to look for a family of segmentations, taking the form of a sequence of partitions $\{P\}_{k\in R^+}$, of decreasing fineness with respect to k.

Taking inspiration from this work, we try here to solve a complementary problem : to evaluate the quality of segmentation results. Starting from the energetic expression of Eq. 1, we propose to characterize segmentation P by the affine and increasing function :

$k \longmapsto E(k, P) = E_D(P) + k \cdot E_C(P)$

3. COMPARISON OF ENERGIES

Regions can be modeled by different functions, the internal energy — or goodness-of-fit energy — is measured by a distance between this model and the set of pixels making up the region. Let R_i be a region containing A_i pixels noted $(X_1, X_2, ..., X_{A_i})$ and let X_p^j be the j-th colour component of pixel X_p. Let μ^j be the average value of component j and let V be the variance / covariance matrix of X_p whose general term is : $V(j,k) = \frac{1}{A_i} \sum_{p=1}^{A_i} \left(X_p^j - \mu^j\right)\left(X_p^k - \mu^k\right)$. Let λ_j be the j-th eigenvalue of matrix V.

The first model is piecewise constant [6] and the distance to the initial data, i.e. internal energy E_D is measured with L_2 norm. For each region, the vector which minimizes this distance is the mean and the distance is : $Q(R_i) = A_i \cdot Trace(V) \quad = A_i \cdot \sum_j \lambda_j$

The second model is a probabilistic model, which supposes that the colour pixels of the region are i.i.d. samples of a Gaussian law. The energy is then the opposite of the log-likelihood of the samples knowing the model. The optimal estimators for mean and variance of a Gaussian law are the empirical estimators and the energy of a region is (except for a constant) [5], provided that for every j, $\lambda_j > 1$:

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

Another form of internal energy very close to these two can also be used : $D(R_i) = A_i \det V = A_i \prod_{j=1}^{3} \lambda_j$

The normalisation coefficient depends on the energy forms. For images coded on $2n$ values by component, n^2 is an upper-bound of the values of variance and covariance. Consequently normalisation is obtained by dividing by $n^2 \times 3 \times A \times 100$ for energy Q, by $n^6 \times 3 \times A \times 10000$ for energy D and by A for energy G.

Concerning the complexity of a segmentation, we simply take the total length of the edges as in the Mumford and Shah model [6]. Normalisation is ensured by a division by the total number of pixels.

We compared these criteria on several results of segmentation, obtained by various algorithms. We display below a comparison on image House, which includes textured and non textured parts and for which visual segmentation is relatively easy and can be consensual. We disposed of 6 segmentation results (cf. Fig. 1) respectively obtained by split and merge algorithm (SM), Tominaga (T), competitive learning (C), region growing (G), 2D histogram classification (H) used in [7] and a fuzzy method (F) [8]. Visually, SM is not accurately segmented since blocks are visible. F is not accurate either, since edges are not very straight, nevertheless the region number for F approaches better the visual perception than the other results (see Table 1). T, C and H include many tiny regions and 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.

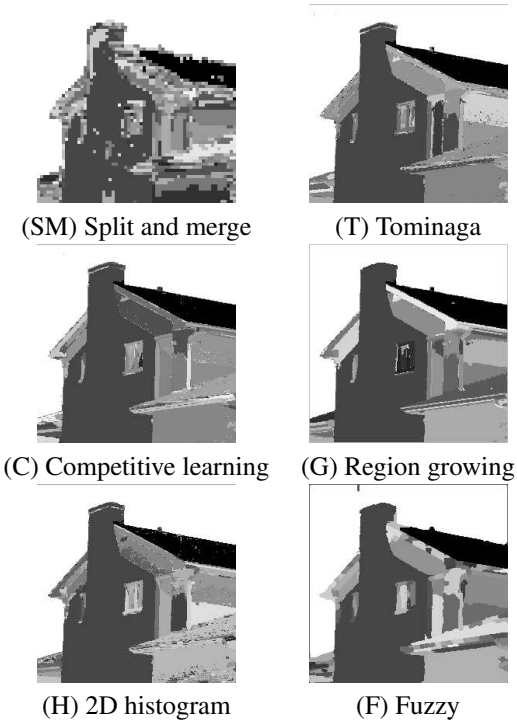


Fig. 1. 6 segmentation results of image House

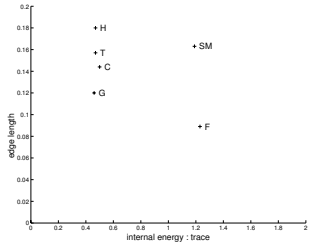
Therefore, the first impression tends to prefer G, which has a legible partition, but a finer examination reveals the spurious tiny regions, favouring F as an alternative good segmentation.

We first computed separately both terms of energy (internal and complexity) and we compared the various forms of internal energy. The aim is also to check if the use of energy is conformable to our visual perception.

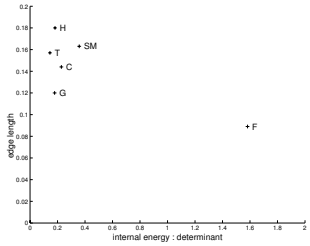
If we represent the couples (E_D, E_C) obtained for the six segmentations (Fig. 2 a,b,c), one can observe that SM is

well discriminated by energy Q, and less well by the other forms of internal energy. Fuzzy method differs from the other results for energies Q and D, as well as for the complexity energy.

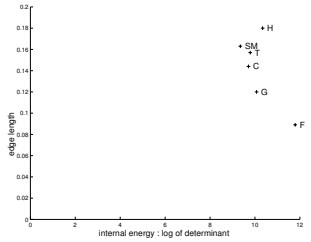
It seems from many results that energy Q provides results which correspond the best to our visual appreciation of the segmented images. As we said previously, parameter k of Eq. (1) is linked to the resolution or the expected degree of precision of the segmentation. Algorithms giving few regions are favored for the complexity energy and disadvantaged for the goodness-of-fit energy.



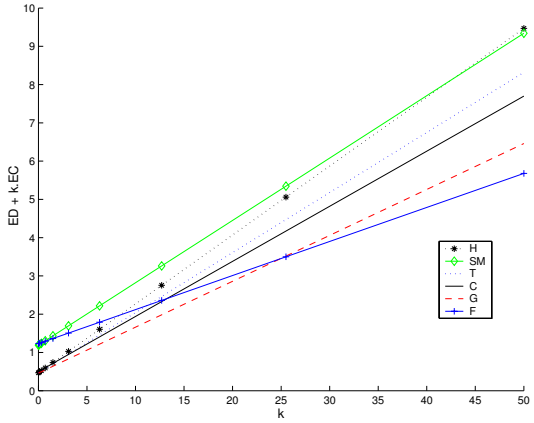
(a) Internal energy : Q



(b) Internal energy : D



(c) Internal energy : G



(d) $E(k, P)$ versus k with internal energy Q (larger k, coarser the resolution)

Fig. 2. Edge energy versus internal energy for 6 segmentation results of image House.

In Fig. 2d, function $k \longmapsto E(k, P)$ is drawn for each partition P of Fig. 2. For all k, the curve representing $E(k, P)$ is always lower for G than for C, T, SM and H. Hence G is always better than these four methods, whatever the scale (or level of detail). The comparison of F and G depends on the expected level of resolution : for a coarse segmentation, F is better than G, and conversely for a fine segmentation. From this graphics, one can conclude that, among the 6 segmentation results, if we look for a coarse segmentation of image House, F gives the best segmentation, and for finer resolutions, G gives the best result.

4. COMPARISON OF EVALUATION CRITERIA

We have compared the main criteria of segmentation evaluation : Levine and Nazif [2], Liu and Yang [3], Borsotti [4] and our energy criterion on a set of images, from which House is very representative. We give the results for the 6 segmentations of image House (cf. Table 1), with two levels of resolution for the energetic criterion ($k = 10$ and 100).

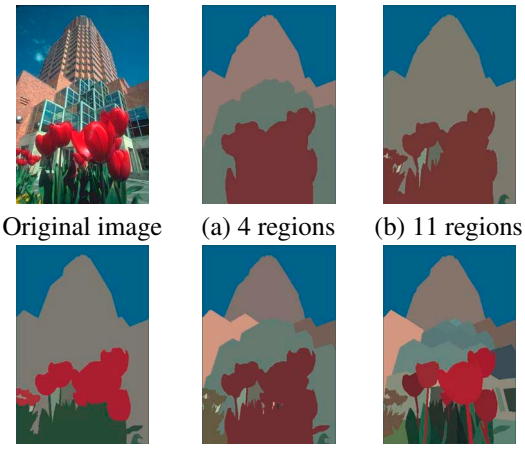
	SM	T	C	G	H	F
Number of regions	379	968	667	379	994	97
Levine-Nazif	116	78	65	49	70	**31**
Liu-Yang	3.2	0.40	0.37	**0.25**	0.39	0.47
Borsotti	0.4	29	8	1.1	24	**0.1**
Energy $k = 10$	2.82	2.04	1.94	**1.66**	2.28	2.12
Energy $k = 100$	17.5	16.2	14.9	12.5	18.5	**10.2**

Table 1 : Comparison of criteria for the 6 segmentation results of image House
In bold, the best result according to the evaluation criterion

All these criteria give no information by themselves, since they are not normalized. They are only useful to compare segmentation results between each other, the best segmentation obtaining the smaller value.

The criterion of Borsotti is extremely sensitive to the number of small regions (one or two pixels). It ranks as first result F as Levine and Nazif does. As expected from Fig. 2, the energy criterion ranks as first result G for a fine resolution ($k = 10$) and result F for a coarse resolution ($k = 100$).

It is interesting to see the behaviour of the tested criteria on images for which we have a ground-truth. We show below the results on an image from the Berkeley database [1] for which 5 manual segmentations are available (Fig. 3). These segmentations have different levels of detail, the number of regions varies from 4 to 67 (cf. Table 2). In Fig. 4 the curve representing (c) is always above the one representing (d), so we can conclude that (c) is better than (d) whatever the resolution, although (d) contains more regions than (c). If a coarse segmentation (large values of k) is expected, it is better to choose (a), for a finer resolution (c) is the best and for the finest resolution (e) is the best. This is in accordance with our visual perception of the results. The other criteria (cf. Table 2) rank results exactly in the order of the region number.



Original image (a) 4 regions (b) 11 regions

(c) 18 regions (d) 27 regions (e) 67 regions

Fig. 3. 5 manual segmentations of an image of the Berkeley database.

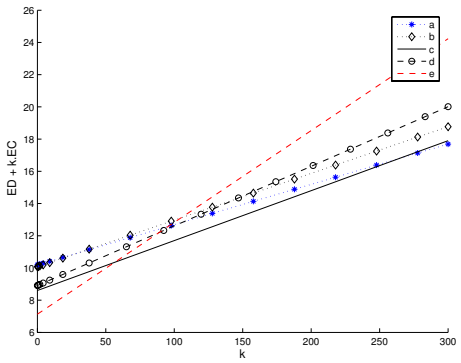


Fig. 4. Variation of E_k versus k with internal energy Q for the 5 manual segmentations of Fig. 3.

	a	b	c	d	e
Number of regions	4	11	18	27	67
Levine and 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
Energy $k = 10$	10.45	10.36	8.91	9.29	**7.70**
Energy $k = 100$	12.7	13	**11.73**	12.65	12.8

Table 2 : Comparison of criteria for the 5 manual segmentations of Fig. 3

5. CONCLUSION

Most existing criteria for segmentation evaluation take into account the distances of the pixels of a region with the average colour of the region and some try to model the complexity of the segmented image by the number of regions. We claim that it is necessary to evaluate a segmentation with respect to the purpose, whose expected level of detail is one of the quantifiable elements. This is why we have proposed criteria of evaluation linked to the level of detail and which take into account both the complexity of the segmentation and the adequation of extracted regions to the original image. The first aspect is measured by the total edge length, which allows to quantify both the number of regions (linked to the resolution) and the regularity of the edges. The second aspect is the goodness-of-fit, quantified for a Gaussian model by a very simple expression. After comparison, it seems that the most efficient amongst these criteria is the Mumford and Shah criterion. As for the complexity of the segmentation, it can be measured by more sophisticated criteria than the edge length, such as the contour regularity, for example.

Moreover, these new criteria are very easy to compute on any kind of image, monochrome or multispectral, with or without edges between regions.

6. REFERENCES

[1] D. R. Martin, *An empirical approach to grouping and segmentation*, Ph.D. thesis, University of California, Berkeley, USA, 2002.

[2] M.D. Levine and A.M. Nazif, "Dynamic measurement of computer generated image segmentations," *IEEE Trans. on PAMI*, vol. 7, no. 25, pp. 155–164, 1985.

[3] J. Liu and Y.-H. Yang, "Multiresolution color image segmentation," *IEEE Trans. on PAMI*, vol. 16, no. 7, pp. 689–700, 1994.

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

[5] L. Guigues, H. Le Men, and J.-P. Cocquerez, "Scale-sets image analysis," in *Proc. of IEEE Int. Conf. on Image Processing (ICIP'03), Barcelona, Spain*, September 2003.

[6] 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.

[7] A. Trémeau, C. Fernandez-Maloigne, and P. Bonton, *Image numérique couleur*, Dunod, Paris, 2004.

[8] 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.