PERGAMON

Pattern Recognition 34 (2001) 969–980

PATTERN RECOGNITION
THE JOURNAL OF THE PATTERN RECOGNITION SOCIETY

www.elsevier.com/locate/patcog

A measure of quality for evaluating methods of segmentation and edge detection☆

Ramón Román-Roldán[a,*], Juan Francisco Gómez-Lopera[a], Chakir Atae-Allah[a], José Martínez-Aroza[b], Pedro Luis Luque-Escamilla[c]

[a]*Dept. Física Aplicada, Universidad de Granada, Campus Fuente Nueva, 18071 Granada, Spain*
[b]*Dept. Matemática Aplicada, Universidad de Granada, Campus Fuente Nueva, 18071 Granada, Spain*
[c]*Dept. Ingeniería Mecánica y Minera, Universidad de Jaén, Alfonso X el Sabio, 28. 23700, Linares, Jaén, Spain*

Received 7 August 1998; received in revised form 27 January 2000; accepted 27 January 2000

Abstract

A new measure of quality is proposed for evaluating the performance of available methods of image segmentation and edge detection. The technique is intended for the evaluation of low error results and features an objective assessment of discrepancy with respect to the theoretical edge, in tandem with subjective visual evaluation using both the neighbourhood and error-interaction criteria. The proposed mathematical model is extremely simple, even from the perspective of computational execution. A training of the measure has been put in practice, which uses visual evaluation of a set of error patterns by a team of observers. Encouraging results were obtained for a selection of test images, especially in relation to other recently proposed and/or currently employed quality measures. © 2001 Pattern Recognition Society. Published by Elsevier Science Ltd. All rights reserved.

Keywords: Performance evaluation; Quality measure; Edge detection; Image segmentation; Subjective assessment of edge detection

1. Introduction

1.1. Quality measures for edge detection and segmentation methods

Authors currently working in the field of low-level image segmentation frequently point to the need for a standard quality measure that would allow the evaluation and comparison of the variety of procedures available [1,2]. Some researchers [3–6] have even stepped outside their usual sphere of interest in an effort to satisfy this demand by proposing new quality measures, as indeed is true of the present authors. Several recent review articles have also revealed the lack of agreement about such a measure.

Take, for instance, the *Dialogue* published in CVGIP: Image Understanding, pp. 245–265, a review of papers carried out by Haralyck [7]. Above all, this study exposed the need for a unified, protocolised assessment of the performance of computer vision procedures, along with a final judgement of results. The dialogue encourages the scientific community working in the field of computer vision to agree to a fixed measure of performance so that the algorithms may be compared both by task-effectiveness and human appreciation. We share the opinion of these authors that in a low-level procedure (e.g., for edge detection) noise immunity is the main feature to be considered for performance characterisation.

Due to subjectivity inherent to the quality concept itself, human intervention is needed to establish certain criteria and/or assessments. There are recent precedents of this practice, like the work of Chalana et al. [8], in which the shape and position of the boundary in medical

☆This work was supported in part by grant TIC94-535 from the Spanish government.

* Corresponding author. Tel.: + 34-958-244-161; fax: + 34-958-243-214.

E-mail addresses: rroman@ugr.es (R. Román-Roldán), jfgomez@ugr.es (J.F. Gómez-Lopera), jmaroza@ugr.es (J. Martínez-Aroza), peter@ujaen.es (P.L. Luque-Escamilla).

PII: S0031-3203(00)00052-2

images is defined by means of a consensual agreement of observers.

Although the problem is often referred to as one of determining the quality of a segmentation method, most of the algorithms for computing the proposed measures work within a window sliding across the image, thus whether or not the image has been effectively segmented is not taken into account. Only a few methods actually explore the obtained segments [9], most measures being best suited to edge detection. Yet in the design of segmentation quality measures, effective edge detection can prove to be highly useful. The present proposal is based on this low-level framework.

1.2. On existing quality measures

Recent surveys [4–6,9] have revealed the shortcomings of currently available measures. Here we briefly describe the four most frequently used and point out their drawbacks.

The discrepancy between a binary image of estimated borders and the ideal one can be stated by means of some kind of account of individual differences, pixel by pixel, between these two binary images. The following terminology and notations are used:

Mistake the discrepancy between the detected edge and the real edge is due to individual discrepancies arising from pixel to pixel, which are here referred to as mistakes. These may be of two kinds:
Bit a mistake due to excess, when a pixel is erroneously defined as an edge pixel.
Hole a mistake due to failure of the method to recognise an edge pixel as such.
N_b no. of bits in the real segmented image.
N_h no. of holes in the real segmented image.
N_e no. of edge pixels in the ideal segmented image.
N no. of pixels in the image.

Once the notation is established, the analysis of common measures is clearer.

1.2.1. Error probability, P_e

$$P_e = \frac{N_b}{N_e}.$$

This measure was proposed by Peli [10]. P_e varies inversely with respect to the quality of the result and is not very useful because holes errors are not accounted for.

1.2.2. Discrepancy D

A discrepancy measure has recently been proposed by Lee [11] for the evaluation of segmentation. For the classic object-background problem — classification into two unique categories — the measure is defined as

$$D = P(O)P(B|O) + P(B)P(O|B),$$

where $P(O)$ is the prior probability of a pixel being classified as belonging to an object, $P(O|B)$ the probability of a background pixel being classified as belonging to an object, $P(B)$ the prior probability of a pixel being classified as belonging to the background, and $P(B|O)$ the probability of a pixel from an object being considered as belonging to the background.

According to Zhang [4], this quality measure is one of the best. In the present case, the objects are the edges of the image, and the remaining pixels belong to the background. Thus,

$$P(O) = \frac{N_e}{N}, \quad P(B|O) = \frac{N_h}{N_e},$$

$$P(B) = \frac{N - N_e}{N}, \quad P(O|B) = \frac{N_b}{N - N_e}$$

and the discrepancy is

$$D = \frac{N_e}{N}\frac{N_h}{N_e} + \frac{N - N_e}{N}\frac{N_b}{N - N_e} = \frac{N_b + N_h}{N}.$$

This measure is clearly better than P_e, since both types of error do appear in the numerator. However, its response is poor, since the measure of an isolated error is independent of its distance from the theoretical edge.

1.2.3. Merit factor, f

This measure was proposed by Pratt [2], and has been greatly utilised for comparing different segmentation methods. It is defined as follows:

$$f = \frac{1}{M}\sum_{i=1}^{N_e - N_h + N_b}\frac{1}{1 + \alpha\, d(i)^2},$$

$$M = \max(N_e, N_e - N_h + N_b),$$

where α is a scale factor (normally $\alpha = 1$), and $d(i)$ is the distance of each of the pixels marked as edge to the theoretical edge. The measure is normalised ($f \in [0, 1]$) and increases with the quality of the segmentation, thus $f = 1$ means a perfect segmentation. Although better than the preceding measure, this does not give a good response in general. Indeed, let us suppose $N_h > N_b > 0$. In this case

$$M = \max(N_e, N_e - N_h + N_b) = N_e,$$

$$\Rightarrow f = \frac{1}{N_e}\sum_{i=1}^{N_e - N_h + N_b}\frac{1}{1 + \alpha\, d(i)^2}$$

and in these conditions an increase in N_h implies a rise in f.

1.2.4. Expanded merit factor, f_e

It is an improved version of f for dealing with low error segmented images. It is defined as [4]

$$f_e = \begin{cases} \frac{1}{N_b} \sum_{i=1}^{N_b} \frac{1}{1+\alpha\, d(i)^2} & \text{if } N_b > 0, \\ 1 & \text{if } N_b = 0. \end{cases}$$

This measure is normalised in $[0, 1]$ and its increasing rises with the quality. It has the disadvantages of not being additive and taking into account positive errors only.

1.3. The framework of the proposed measure

The proposed measure is a hybrid of the so-called "*empirical discrepancy*" and "*empirical goodness*" by Zhang [4]. It combines the pixel-by-pixel objective discrepancy between the obtained edges and the theoretical ones with an evaluation of each mistake as assessed by human observers.

The measure belongs to a class that may be called the *low error model* of quality measures, not intended for excessive or aberrant (not usually found in edge images) errors. In other words, the measure should be properly applied to edge images with no too many mistakes, and most of these should be close to the theoretical edge.

2. Characterizing the measure

In order to implement the initial statement (discrepancy, empirical goodness and low error) in a mathematical expression for a quality measure R, the following bases are established.

2.1. Discrepancy

The theoretical edge must be known so that the mistakes produced can be detected and evaluated one by one. However, the edge must be defined by convention. When two homogeneous regions are in touch and the boundary has to be defined as a one-pixel-thick dividing line, the newly defined edge pixels clearly must be placed in one of the two areas. Two different methods might perfectly locate the desired boundary within an image, but while one method will draw the edge on the darker side, the other will always place it in the lowermost region. The binary images of the edges obtained by the two methods will obviously be different, but both should be classified as error-free since the actual detection was perfect. To prevent this ambiguity in edge definition from influencing the number of mistakes, the theoretical edge may be assumed to lie on the side with the fewest mistakes.

The measure R should be expressed as the sum of the values assigned to all mistakes in the edge image. These values must take into account the presence (or not) of other mistakes produced close to the one being evaluated. Therefore, a sum expression for R does not imply that it be a merely additive function of single mistakes, nor even an increasing function with respect to the number of mistakes.

2.2. Empirical goodness

This feature is based on two criteria for the appearance of a mistake in the edge image.

2.2.1. Neighbourhood

The way a human observer appreciates an edge image depends on the distance of each mistake from what is perceived to be the true edge. However, this is true only for very short distances. In fact, a mistake (bit) produced just next to the edge is assumed to be linked to it, while bits located only two or three pixels away from the edge are no longer perceived thus. Hence in the proposed measure, by contrast with some quality measures which include the distance to the edge as a variable, a mistake value will not depend on distance if this is greater than one or two pixels. The fact that only strict closeness to the edge will be taken into account is then an essential and novel criterion that introduces a notable simplification in the mathematical expression for R.

2.2.2. Mistake interaction

Also, human appreciation of closely clustered mistakes is more acute than when the same mistakes are scattered throughout the image. An observer assessing a group of mistakes lying within a small neighbourhood (a window W) will allot them different values (interaction), and these may be extenuating or aggravating.

For W centred on a given mistake, if there is one or more mistakes of the same type then visual appreciation is generally more severe. Conversely, if the accompanying mistakes are not of the same type and are placed horizontally or vertically, not obliquely, it will appear that the true edge pixels have been only slightly shifted, and the connectivity of the edge is preserved. Hence, the evaluation must be extenuating for mistakes of different type in W, but aggravating for the same type. The window W must be chosen small, since visual appreciation claims for short-range aggravating and extenuating interactions.

2.3. A low-error model

The low error nature of the proposed method means that the evaluation of empirical goodness will be performed within a small window centred on each mistake. This, jointly with the short-range interactions, leads us to choose the simplest solution: a 3×3 square window for individual error evaluation. In practice, the results

justified this decision and any improvement gained using larger window sizes would not appear to merit the extra effort required.

The low error model assumed for this method only exclude very bad, aberrant detection edge algorithms. However, some results are given for a few extra noisy, out of model segmented images included in the series of Fig. 5 (the sixth of each).

3. The measure of quality

Following these premises, the construction of the mixed measure of quality R is now summarised by introducing an adjustable parameter accompanying to each feature to be taken into account.

1. Because of the focusing of R for low error rate, it is primarily defined as the number of (independent) mistakes in the image. In order to allow for different assessment of holes and bits and to scale the measure, two coefficients are introduced: $R = aN_b + cN_h$.
2. The number of mistakes are spanned in sums for assessing each mistake individually: $R = a\sum_{bits} e_b + c\sum_{holes} e_h$, where the bit and hole errors are given the plain values $e_b = e_h = 1$ initially.
3. Aggravating interaction between same type mistakes is now taken into account by developing the error summands as $e_b = 1 + bn_b$, $e_h = 1 + hn_h$, where b, h are coefficients to be adjusted, and n_b is the no. of bits in W, minus the central mistake, n_h the no. of holes in W, minus the central mistake.
4. Extenuating interaction between different type mistakes is introduced by means of a dividing factor

$$e_b = \frac{1 + bn_b}{1 + i_{bh}n'_h}, \quad e_h = \frac{1 + hn_h}{1 + i_{hb}n'_b},$$

where i_{bh}, i_{hb} are new coefficients to be adjusted, and n'_b the no. of bits in direct contact with the central mistake, n'_h the no. of holes in direct contact with the central mistake.

The quotient format has been selected to get e_b, e_h positive and decreasingly sensible with n'_h, n'_b, respectively. (Obviously, multiplicative factors like $(1 - i_{bh}n'_h)$ do not have these desirable properties.)
5. Finally, the neighbourhood criterion for bits (no place for holes neighbourhood) is now taken into account by adding a new term in the denominator of e_b (the same rationale than the extenuating interaction apply). Being p the proximity to the edge coefficient and n_e the number of real edge pixels in W, the definitive expression for R is

$$R = a\sum_{bits} \frac{(1 + bn_b)}{1 + pn_e + i_{bh}n'_h} + c\sum_{holes} \frac{(1 + hn_h)}{1 + i_{hb}n'_b}$$

4. Determining parameters

In order to determine accurately the coefficients of R, a training procedure has been established, consisting of:

1. Selecting a set of *error patterns*.
2. Assigning them numerical values, based on subjective human assessment provided by a team of observers.
3. Approximating R as close as possible to the agreed values by adjusting the coefficients involved in the formula for R, based on the reasonable assumption that R will henceforth produce acceptable values for real segmented images.

4.1. The set of training error patterns

A set of individual and small groups of errors were selected, according to the following criteria:

- They should be viewed as *typical* for edge detection, in accordance with the initial specification, which excludes application of R to aberrant errors.
- The size and shape of the patterns were to be selected from a linear edge, taking into account the previous requisite.
- Patterns were to be varied as possible, while not excessively numerous, for obvious reasons.

Fig. 1 shows the set of patterns used to adjust the coefficients. Series A is for certain non-straight configurations, series B contains only holes, series C only bits, and series D is mixed.

4.2. Subjective assessment of error patterns

Each pattern was assigned a value representing the visual impression of the observer. This is, in fact, the only possible way to achieve a good quality measure in accordance with the empirical-goodness criterion. Nevertheless, it should be borne in mind that observers were image processing experts and could not contract out of evaluating factors such as ease of restoration, edge connectivity, or absence of bits far from the theoretic edge. Of course, the aforementioned criteria for constructing the quality measure were employed at the time of assessment, since such a measure should not behave other than as proposed. Great care was thus taken — as the success of adjustment depends on it — over dissociating visual impressions obtained while gazing at a pattern from the mathematical expression with which the pattern would then be assessed.

The team of observers performing the subjective assessment included the eight members of the research group to which the present authors belong. The following

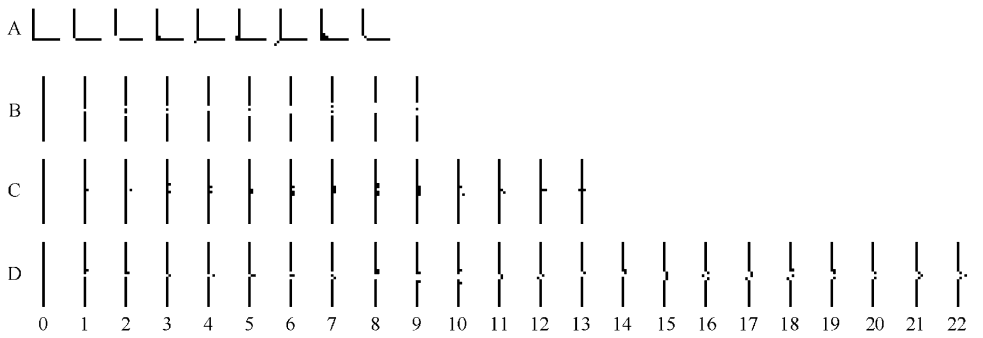

Fig. 1. Set of patterns for subjective evaluation.

assumptions were made as a starting point:

- The quality scale begins at zero, representing absence of error, and increases with severity of error. Thus it is really a scale of badness rather than of quality.
- By convention, the unit of the scale represents a single hole (see pattern B_1).
- No upper bound is stated for badness (the scale is not normalised). This is because the measure is designed to compare the performance of different segmentation methods on the *same set* of images, thus obviating normalisation.
- The value assigned to a pattern should correspond to a subjective impression. However, since each study includes a great number of successive comparisons between only two or three patterns at any one time, it is again important to rationalise the set of values in order to prevent the occurrence of inconsistent or even contradictory pairs.

Finally, an overall value is agreed upon, or an average is taken in case of disparity. The final subjective values thus obtained are given in Fig. 2(a) after proper adjustment.

4.3. Adjusting coefficients of R

The adjustment of coefficients of R has been carried out by the least-squares method with respect to the subjective values. The mean square error (*RMSE*) is to be minimised for the optimum adjusted R; besides, it permit to compare different quality measures for the same set of patterns. No further normalisation of R is needed for this purpose. The adjusted coefficients by the above procedure are shown in Table 1.

Results are plotted in Fig. 2. The X-axis shows the training patterns, while the Y-axis represents the subjective values assigned to errors (curve a) and values obtained by R (curve b). They have been ordered by their increasing subjective values. Due to the number of coefficients in the expression, the measure R is a very versatile one, widely adaptable to other reference patterns.

Table 1
Numerical values of coefficients in R

	R	R modified by Euler's characteristic
a	2.02189276	1.94447565
c	1.70510940	1.73564803
b	0.015966617	0.01311035
p	0.166866567	0.14738691
i_{bh}	12.38602179	4.51205205
h	0.414879829	0.37368425
i_{hb}	0.144839388	0.08645188
c_{Euler}	—	8.92970173

4.4. Response of other quality measures to the training patterns

Although error patterns were selected and assessed with the aim of determining R, a comparison of the present results with those of other quality measures allows us to obtain a preliminary estimation of the characteristics of the compared measures, always from the point of view of the initial approach to the problem — discrepancy, empirical goodness and low error. To perform this comparison, the following measures from the literature were applied to the same error patterns: merit factor f, expanded merit factor f_e, discrepancy D, and error probability P_e.

To facilitate comparison, results for D, P_e, f and f_e are shown in Fig. 3, set out similarly to Fig. 2. Measures f and f_e have been inverted, and all of them have been adjusted by minimum square with respect to the subjective values.

Some comments are worthy:

- Measures P_e, f and f_e give non-sensical results with respect to the agreed values.
- Measure D has a relatively good behaviour due to its additivity, which is coincident with the additivity inherent to the subjective assignations.

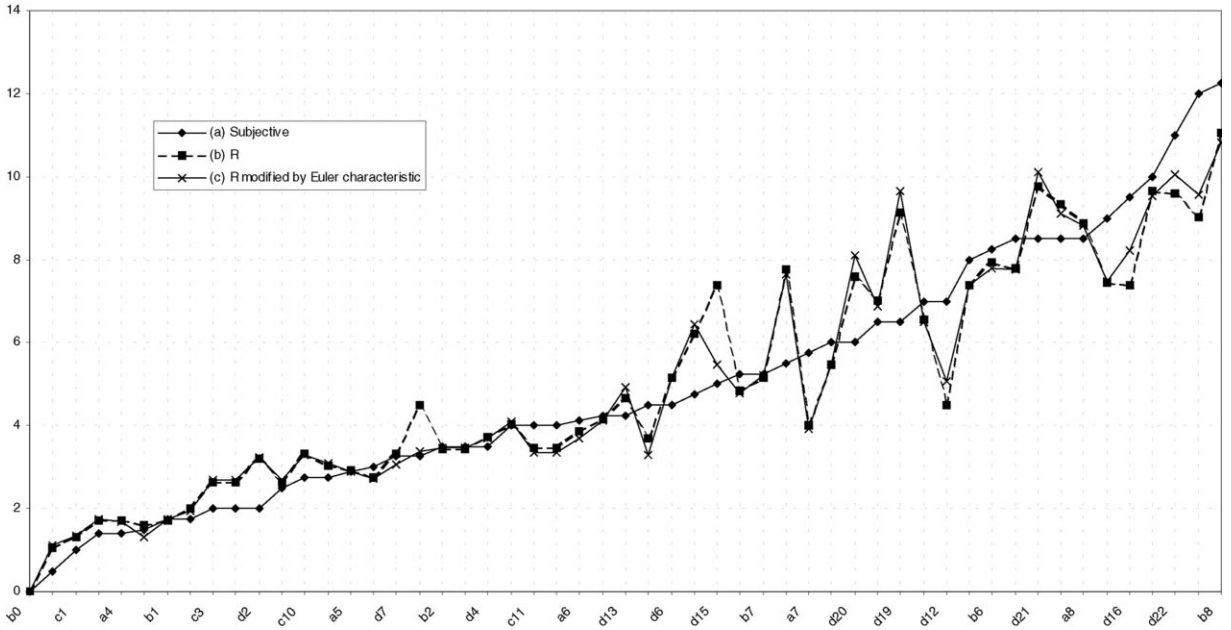

Fig. 2. Normalised pattern values for (a) subjective, (b) R, (c) R modified by Euler's characteristic.

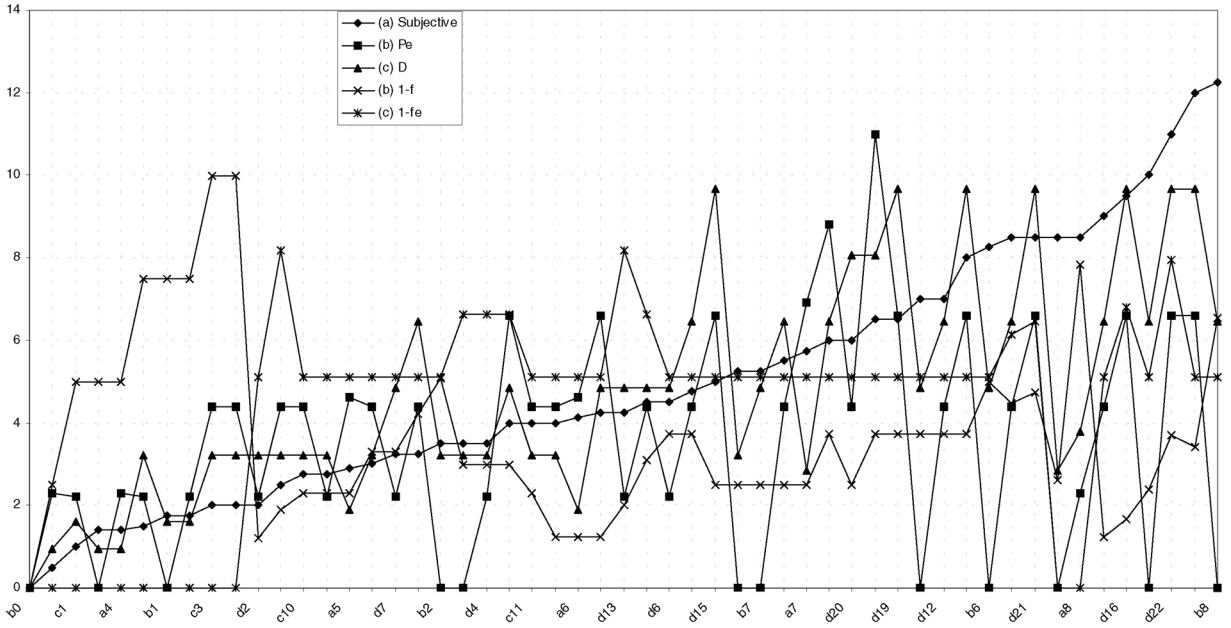

Fig. 3. Normalised pattern values for (a) subjective, (b) P_e, (c) D, (d) $1-f$, (e) $1-f_e$.

It may be observed that the proposed quality measure R fits especially well to the subjective values, when compared against other measures. There are three patterns which produce a not-so-good fit: D_{12}, D_{16} and D_{18}. These are patterns deliberately intended to produce a failure of R, and are not very typical — as was indeed intended from the outset. In accordance with basic premises, a highly displaced edge is penalised but without further attenuation even when there are holes nearby. To deal with these special cases, R is corrected by means of Euler's characteristic.

Table 2
Numerical results of RMSE obtained for several quality measures with respect to the subjective values

Quality measure	*RMSE*
R	1.084958236
R + Euler's	1.014406947
P_e	3.789961359
D	2.064971497
$1-f$	3.994423645
$1-f_e$	3.023500544

The RMSE values capture the global different behaviour of the compared measures. They also confirm the better performance of R, as shown in Table 2.

4.5. A further improvement: Euler's characteristic for connectivity

Some patterns are not so well assessed by R, probably due to a lack of connectivity not explicitly included in the mathematical expression (see patterns D_{12}, D_{16} and D_{18}). This deficiency can be partially repaired by introducing a new coefficient in R, which takes this fact into account. The topological coefficient of Euler (G) was chosen here, resulting in the enhanced performance shown in Fig. 2, curve (c).

Given a pattern of pixels in a binary image, it is possible to define its *characteristic of Euler*. This is a topological invariant including information about the connectivity of the pattern [12]. It is defined as

$$G = V - E + P$$

$$\text{where} \begin{Bmatrix} V = \text{no. of vertices} \\ E = \text{no. of edges} \\ P = \text{no. of pixels} \end{Bmatrix} \text{in the pattern.}$$

Objects shared by more than one pixel are accounted for only once in the above expression. Fig. 4(a) shows a pixel (in grey) with its four vertices and its four edges.

G has the property of being 1 if the pattern is connected, that is, if every pixel in the pattern has a neighbour (by an edge or by a vertex) in the pattern. However, the converse is not always true. As an example, Fig. 4(b)–(d) shows three patterns. The corresponding values of the Euler's characteristic are $G_{4(b)} = 8 - 10 + 3 = 1$ (connected), $G_{4(c)} = 8 - 8 + 2 = 2 \neq 1$ (not connected) and, $G_{4(d)} = 4 - 4 + 1 = 1$ (not connected). The influence of G in R takes place in the denominator of the part of R which evaluates holes, since the lack of connectivity comes from holes. For it, a new coefficient c_{Euler} is introduced in such a way that, for holes keeping edge connectivity, R is decreased ($c_{Euler} > 1$). The connectivity close to a hole pattern in W is considered as preserved if there is a bit in touch with each hole within W:

$$R = a\sum_{bits} \frac{(1 + bn_b)}{1 + pn_e + i_{bh}n'_h} + c\sum_{holes} \frac{(1 + hn_h)}{1 + c_{Euler}i_{hb}n'_b},$$

$c_{Euler} > 1$ if $G = 1$ and $\forall$ hole in W the central pixel has a bit in direct contact, $c_{Euler} = 1$ in other case

The enhancement obtained through G can be seen by comparing D_{11} and D_{12}. Before making the correction, D_{12} was evaluated better than D_{11}, against visual impression. This is because D_{11} has an additional interaction between bits that D_{12} does not have. After making the correction with G, D_{12} has a value higher than D_{11}, thus agreeing with the visual impression. Although pattern a_2 (very seldom in practice) is contrarily valued, the inclusion of G generally improves the performance or R, as can be seen in Fig. 2. Table 1 shows numerical values of the parameters in R modified by Euler characteristic obtained by least-squares adjustment. The RMSE value for the G-modified R is therefore lower than for the unmodified R (see Table 2).

4.6. Sensitivity of R with respect to the coefficients

To assess the confidence in the adjustment of the coefficients, the sensitivity of R with respect to them must be considered. Since a theoretical treatment is cumbersome, some representative, numerical results have been obtained.

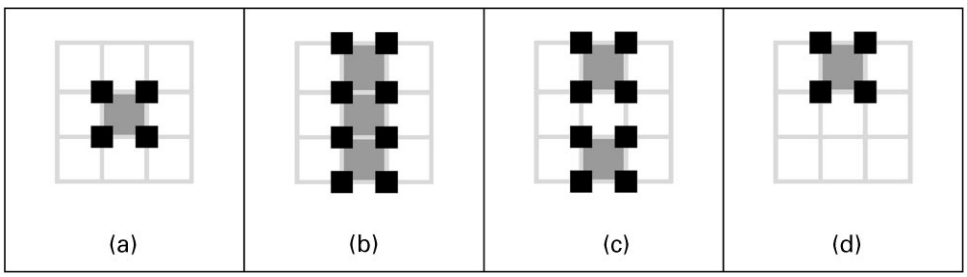

Fig. 4. (a) Edges and vertices of a pixel, (b) connected set of pixels, (c) unconnected set of pixels, (d) unconnected hole with $G = 1$.

Table 3
Sensitivity of R to variations of coefficients. Error (%) in *RMSE* provoked by several variations (%) of each coefficient from its optimal value

	+10%	−10%	+50%	−50%
W	2.65	2.65	53.1	53.1
B	5.27×10^{-4}	5.27×10^{-4}	1.32×10^{-2}	1.32×10^{-2}
P	0.160	0.181	3.16	5.77
i_{bh}	3.64×10^{-2}	5.02×10^{-2}	0.540	2.91
h	1.04	1.04	23.4	23.4
i_{hb}	6.77×10^{-2}	7.81×10^{-2}	1.31	2.71
c_{Euler}	5.00×10^{-2}	5.95×10^{-2}	0.914	2.20

Table 4
Exact and rounded final values of the R coefficients and their *RMSE*

	Original value	Rounded value
K	1.73564803	1.7
w	1.1203168	1.1
b	0.01311035	0.013
p	0.14738691	0.15
i_{bh}	4.51205205	4.5
h	0.37368425	0.37
i_{hb}	0.08645188	0.086
c_{Euler}	8.929701731	8.9
RMSE	1.0144069	1.0276189

For this purpose, the above expression of R is conveniently modified to the following by substituting the coefficients $a = Kw$ and $c = K$:

$$R = K\left[w\sum_{bits}\frac{(1+bn_b)}{1+pn_e+i_{bh}n'_h} + \sum_{holes}\frac{(1+hn_h)}{1+c_{Euler}i_{hb}n'_b}\right].$$

Now, K is a scale factor, irrelevant for the sensitivity analysis, and w is a bit-hole balance factor. Table 3 shows the RMSE obtained for R when each coefficient has been changed separately + 50, + 10, − 10 and − 50% from their original, optimum settings, as well as the corresponding variations (in %) of *RMSE*.

All relative variations in RMSE are no greater than that of the provoking variations in the coefficients. These results say that the adjusted values of the coefficients are reliable for using R confidently.

A generally symmetric response of the *RMSE* errors makes evident for 10% coefficient variations, as well as for some (w, b, h) at higher variations.

Some *RMSE* errors should be noticed. As expected, the highest values appear for strong variations (50%) in w and h, the bit-hole balance and the aggravating hole interaction, respectively. On the other hand, the errors for b result extremely low. This has the meaning that the aggravating bit interaction is almost irrelevant. It may be due to the fact that the training patterns do not include clusters of bits not close to the edge. For specific applications, such that comparing edge detectors for very noisy images, it could be convenient to use a different set of training patterns, leading to higher values for b and the corresponding sensitivity.

Table 4 contains final values of the coefficients as determined by the least-squares adjustment. The round-off to the nearest two digits value has been made according to the sensitivity results: being 5% the maximum relative error in the coefficients, a little error in RMSE, only 1.3%, is produced.

5. Results and comparison with other quality measures

The proposed measure and other available measures were applied to a selected set of edge images provided by segmentation procedures and results were compared in order to assess the behaviour of R.

5.1. Test images

In order to achieve comparative results about different methods, four series of test edge images were selected. By visual comparison between them, each series was ordered according to quality by consensus of the observers. The images (see Fig. 5) were selected in such a way that, in the opinion of the members of the team, there was no doubt about the correct arrangement. That is, the difference in quality between each two consecutive images in a series was great enough to expect any other qualified observer to agree.

These were the criteria used in selecting the test images:

- They were synthetic images, this being a requirement for the application of the quality measures proposed for this comparison, and most were of the "discrepancy", or mixed, type.
- They were of the same type as those utilised by other published studies on this subject [13] (new proposed measures, or surveys of measures), in order that the reader may more easily assess the results.
- Bearing in mind that R is designed as a hybrid measure, the prior assessment of test images must also be performed using the two criteria, empirical goodness and discrepancy. Since combining both in a subjective appreciation is difficult, they were applied separately. Thus, the first three series contain scattered and grouped mistakes, but no systematic error, whereas part (d) in Table 4 refers to images with only continuous errors — displacements, duplications, etc.

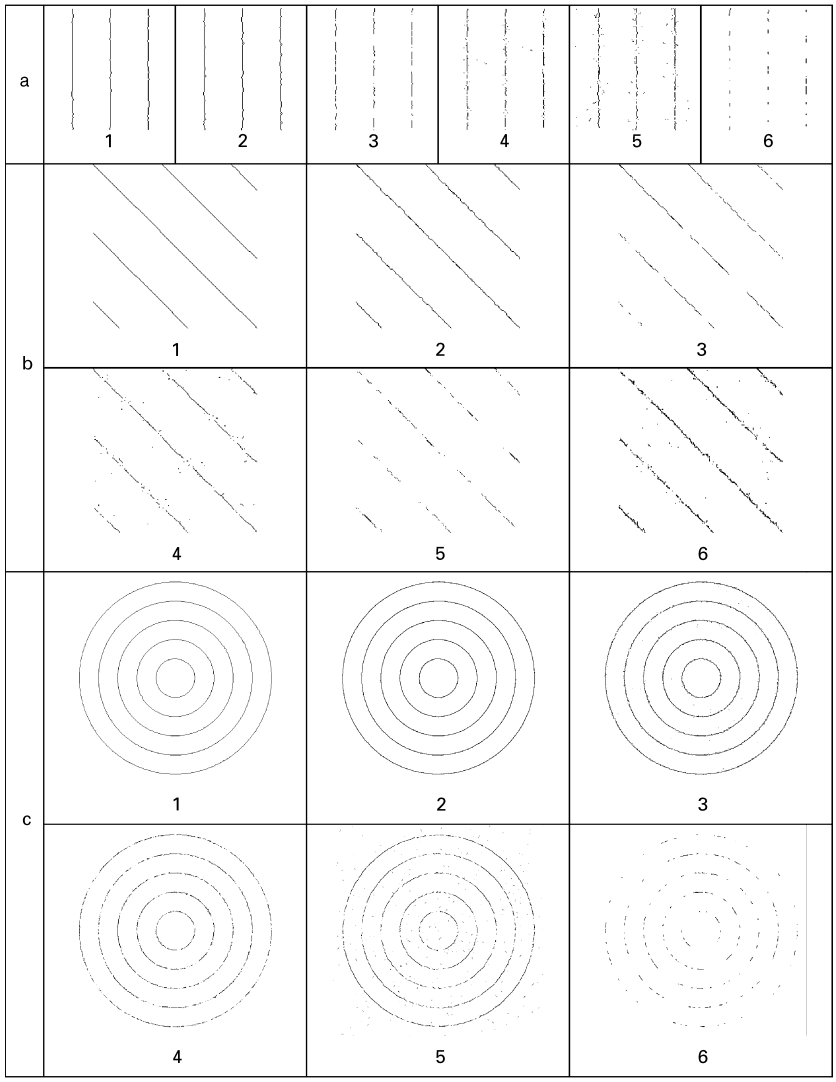

Fig. 5. Series of segmented images.

The fourth series has not been shown in Fig. 5 owing to obviousness. The real edge is a straight vertical line, and the test edges are as follows: (1) a two-pixel-thick vertical line over the edge; (2) a one-pixel-thick vertical line displaced one pixel to the right with respect to the real edge; (3) a three-pixel-thick vertical line centred on the edge; (4) same as (3) but shifted one pixel to the right; (5) same as (1) but shifted one pixel to the right, thus not covering the real edge; and (6) same as (5) but shifted again to the right.

5.2. Order deviation

Besides R, the quality measures P_e, f, D, f_e were applied to the test images. The unnormalised numerical results are given in Table 5. At the right of each set of numerical values, a corresponding set of ordinals indicates how the measure has ordered the test images. The comparison of the results was performed not directly from the numerical values, but from these resulting ordered sequences instead.

In order to assess the level of success on each one of the above-mentioned sequences, it must be stated some kind of distance measure of a sequence $s = (s_1, s_2, \ldots, s_n)$ from the ideal $(1, 2, \ldots, n$ by construction). This measure, $T(s)$, is defined as the number of relative disorders between each two elements of s. For example in the sequence (4, 2, 6, 3, 1, 5) there are nine relative disorders. This is what we wish to determine, instead of the mere deviation

Table 5
Numerical results, ordered sequences and order deviations

Serial	S_c	$10P_e$	S_{pe}	10^2D	S_D	$10f$	S_f	$10f_e$	S_{fe}	$10^{-2}R$	S_R
	1	1.54	3	0.58	1	9.23	1	5.00	1[a]	1.30	1
	2	2.49	4[a]	0.90	2	8.72	2	4.72	4	2.07	2
	3	1.16	2	1.03	3	6.30	5	5.00	2[a]	5.07	3
A	4	2.51	5[a]	1.28	4	6.74	4	3.82	5	5.77	4
	5	4.54	6	1.66	6	6.92	3	2.99	6	6.57	5
	6	0.27	1	1.51	5	2.44	6	5.00	3[a]	10.40	6
	T(*s*)		*7.5*		*1*		*3*		*5.5*		*0*
	1	3.45	1	0.78	1	8.29	1	5.00	1	2.42	1
	2	7.28	5	1.43	3	6.91	2	4.83	2[a]	5.42	2
	3	5.45	3	1.46	4	5.43	5	4.80	3[a]	9.11	3
B	4	6.92	4	1.56	5	5.82	4	3.75	6	9.75	4
	5	3.77	2	1.37	2	3.69	6	4.66	4	12.23	5
	6	12.49	6	1.86	6	5.89	3	3.80	5	13.12	6
	T(*s*)		*5*		*3*		*4*		*2.5*		*0*
	1	5.41	2	1.51	2	7.27	1	4.86	1	46.26	1
	2	8.12	4	1.60	3	7.17	2	4.80	2[a]	56.39	2
	3	10.65	6	2.04	6	6.43	4	4.34	5	79.10	3
C	4	6.06	3	1.68	4	6.81	3	4.59	4	80.56	4
	5	8.31	5	2.00	5	5.94	5	3.88	6	99.43	5
	6	1.72	1	1.43	1	2.41	6	4.79	3[a]	108.80	6
	T(*s*)		*8*		*7*		*1*		*4.5*		*0*
	1	10.0	1[a]	1.52	1	7.50	1	5.00	1[a]	7.86	1
	2	10.0	2[a]	3.03	2[a]	5.00	4	5.00	2[a]	12.02	2
	3	20.0	3[b]	3.03	3[a]	6.67	2	5.00	3[a]	15.72	3
D	4	20.0	4[b]	3.03	4[a]	5.67	3	3.50	4[b]	19.99	5
	5	20.0	5[b]	4.55	5[b]	3.50	5	3.50	5[b]	23.93	4
	6	20.0	6[b]	4.55	6[b]	1.50	6	1.50	6	41.23	6
	T(*s*)		*3.67*		*2*		*2*		*2*		*0*

[a,b]The values in the series are different less than the 1% of the series' range.

of each element from its correct location in the sequence. As a consequence of this definition, $T(s)$ is an integer number between 0 and $\binom{n}{2}$. In addition, $T(s)$ can be alternatively defined by any of the following equivalent statements:

1. $\sum_{k=2}^{n} L(k)$, where $L(k) =$ no. of $\{s_i \in s, s_i > s_k, i < k\}$ is the number of elements greater than the one occupying location k and found to the left.
2. $\sum_{k=1}^{n-1} R(k)$, where $R(k) =$ no. of $\{s_j \in s, s_k > s_j, k < j\}$ is the number of elements smaller than the one occupying location k and found to the right.
3. The minimum number of transpositions between two adjacent elements needed to order the sequence.

This last characterisation for $T(s)$ is a consequence of two facts: first, that the permutations group S_n can be generated by transpositions between adjacent elements $(1\ 2), (2\ 3), \ldots, (n-1\ n)$, where $(i\ j)$ stands for a transposition between positions i and j [14]; and second, that every sequence can be ordered by only making transpositions between adjacent elements in relative disorder, as can be seen from an analysis of the well-known bubble sorting algorithm.

Very close values of quality must be considered as equal. So, for each measure, an inequality tolerance of 1% of the range was established. If, according to this, an ordered sequence s contains equalities, then its deviation $T(s)$ is obtained as the average of $T(s_p)$ for all the sequences s_p differing from the standard only in a permutation of equal-valued elements.

5.3. Results

Below each order sequence s, Table 5 also shows the corresponding order deviation $T(s)$. The correct sequence s_c is given in the second column as a reference for easy checking.

The results reported show an excellent behaviour for the proposed measure R. In particular, R is the only that gives the correct order for all the series.

It can be seen that measures D, P_e, f and f_e have relatively good behaviour in certain series, but not in all, being f the best of them. This corroborates the argument given in the criticism at the beginning of this paper.

6. Conclusions

A new measure of quality for evaluating the performance of edge segmentation methods has been proposed. It is a hybrid of empirical discrepancy and empirical goodness, which appears to be the type of measure the best suited to the general problem. Its mathematical expression is simple and intuitive, also having a high computational efficiency. The parameters have been set up by means of a least-squares training process, in order to obtain a measure having a similar behaviour of the human perception. Once the parameters have been adjusted, the measure has been applied to a select set of edge-images. Then, the results have been compared to those obtained by applying other quality measures in literature to the same set of images. The proposed measure shows a noticeable better behaviour than any of the others.

7. Summary

Authors currently working in the field of low-level image segmentation frequently point to the need for a standard quality measure that would allow the evaluation and comparison of the variety of procedures available. Some researchers have even stepped outside their usual sphere of interest in an effort to satisfy this demand by proposing new quality measures, as indeed is true of the present authors. Several recent review articles have also revealed the lack of agreement about such a measure.

Although the problem is often referred to as one of determining the quality of a segmentation method, most of the algorithms for computing the proposed measures work within a window sliding across the image, thus whether or not the image has been effectively segmented is not taken into account. Only a few methods actually explore the obtained segments, most measures being best suited to edge detection. Yet in the design of segmentation quality measures, effective edge detection can prove to be highly useful. The present proposal is based on this low-level framework.

The discrepancy between a binary image of estimated borders and the ideal one can be stated by means of some kind of account of individual differences, pixel by pixel, between these two binary images. The proposed measure is a hybrid of the so-called "*empirical discrepancy*" and "*empirical goodness*" by Zhang. It combines the pixel-by-pixel objective discrepancy between the obtained edges and the theoretical ones with an evaluation of each mistake as assessed by human observers. The measure belongs to a class that may be called the *low-error model* of quality measures, not intended for excessive or aberrant (not usually found in edge images) errors. In other words, the measure should be properly applied to edge images with no too many mistakes, and most of these should be close to the theoretical edge.

In order to implement these statements in a mathematical expression, the following bases are established: first, the theoretical edge must be known (or defined by convention) so that the mistakes produced can be detected and evaluated one by one; second, only strict closeness to the ideal edge is relevant; and third, interactions between closely clustered mistakes should be involved in the assessment of the quality.

By following these premises, a new measure of quality is proposed for evaluating the performance of available methods of image segmentation and edge detection. The technique is intended for the evaluation of low-error results and features an objective assessment of discrepancy with respect to the theoretical edge, in tandem with subjective visual evaluation using both the neighbourhood and error-interaction criteria. The proposed mathematical model is extremely simple, even from the perspective of computational execution. A training of the measure has been put in practice, which uses visual evaluation of a set of error patterns by a team of observers. Encouraging results were obtained for a selection of test images, especially in relation to other recently proposed and/or currently employed quality measures.

References

[1] I.E. Abdou, W.E. Pratt, Quantitative design and evaluation of enhancement/thresholding edge detectors, Proc. IEEE 67 (5) (1979) 753–763.

[2] W. Pratt, Digital Image Processing, Wiley-Interscience, New York, 1991.

[3] P.L. Palmer, H. Dabis, J. Kittler, A performance measure for boundary detection algorithms, Comput. Vision Image Understanding 63 (3) (1996) 476–494.

[4] Y.J. Zhang, A survey on evaluation methods for image segmentation, Pattern Recognition 29 (8) (1996) 1335–1346.

[5] Q. Zhu, Efficient evaluations of edge connectivity and width uniformity, Image Vision Comput. 14 (1996) 21–34.

[6] Q. Huang, B. Dom, Quantitative methods of evaluating image segmentation, IEEE International Conference on Image Processing, 1995, pp. 53–56.

[7] R.M. Haralick, Performance characterization in computer vision, CVGIP: Image Understanding 60 (2) (1994) 245–249.

[8] V. Chalana, Y. Kim, A methodology for evaluation of boundary detection algorithms on medical images, IEEE Trans. Med. Imaging 16 (5) (1997) 642–652.
[9] Y.J. Zhang, J.J. Gerbrands, Objective and quantitative segmentation evaluation and comparison, Signal Process. 39 (1994) 43–54.
[10] C.K. Leung, F.K. Lam, Performance analysis for a class of iterative image thresholding algorithms, Pattern Recognition 29 (9) (1996) 1523–1530.
[11] T. Peli, D. Malah, A study of edge detection algorithms, Comput. Graphics Image Process. 20 (1982) 1–21.
[12] J. Olszewski, A flexible thinning algorithm allowing parallel, sequential, and distributed application, ACM Trans. on Math. Software 18 (1) (1992) 25–45.
[13] D.J. Park, K.M. Nam, R.H. Park, Edge detection in noisy images based on the co-occurrence matrix, Pattern Recognition 27 (6) (1994) 765–775.
[14] S. MacLane, G. Birkhoff, Algèbre, Solutions Dévelopées des Exercices, Premiere Partie. Gauthier-Villars, Paris, section 6, sample 13, 1972, pp. 99.

About the Author—R. ROMÁN-ROLDÁN is Professor of Applied Physics at the University of Granada (Spain). He is currently researching in two interrelated subjects, both based on the use of an information-theoretic measure, the Jensen-Shannon divergence. *Entropic edge detection in digital images*, and *Complexity analysis in DNA sequences through entropic segmentation*.

About the Author—P.L. LUQUE ESCAMILLA received the B.Sc. and Ph.D. degrees in Physics from the Universidad Complutense (Madrid) and Universidad de Granada, in 1992 and 1996 respectively. He is currently professor in the Dpt. Ingenierías Mecánica y Minera in Universidad de Jaén. His research interests include digital image filtering and segmentation, particularly in practical applications (i.e. fluid dynamics).

About the Author—JUAN FRANCISCO GÓMEZ-LOPERA was born in Granada, Spain, on 3 June 1968. He received the M.S. and Ph.D. degrees in Physic from Granada University, in 1991 and 1995, respectively. He joined the faculty of Experimental Sciences of the Department of Applied Physic in 1992, at the University of Almería. Now he is professor at the University of Granada, since 1996. His research interest includes image filtering, image segmentation and its applications.

About the Author—JOSÉ MARTÍNEZ AROZA was born in Archidona, Spain and received his M.S. and Ph.D. degrees in Universidad de Granada. He teaches numerical analysis in this University, in the Dept. Matemática Aplicada. He is interested in image processing and mathematics, as well as in women, of course.

About the Author—CHAKIR ATAE-ALLAH was born in Tangier, Morocco, on 10 September 1970. He received the M.S. in Physic from Tetuan University, Morocco. Since 1994 he is investigator at Department of Applied Physic in the University of Granada, Spain. His research interest includes edge detection, thinning, linking and its applications.