ELSEVIER

Pattern Recognition Letters 19 (1998) 741–747

Pattern Recognition Letters

Quantitative evaluation of color image segmentation results [1]

M. Borsotti [a], P. Campadelli [a,2], R. Schettini [b,*]

[a] *Dip. di Scienze dell'Informazione, Università degli Studi di Milano, Via Comelico 39, Milano, Italy*

[b] *ITIM, Istituto Tecnologie Informatiche Multimediali, CNR, Consiglio Nazionale delle Ricerche, Via Ampere 56, 20131 Milano, Italy*

Received 9 September 1997; revised 11 February 1998

Abstract

In this paper we consider the problem of the automatic evaluation of the results of color image segmentation. Liu and Yang (1994) have proposed an evaluation function, inspired by the qualitative criteria for good image segmentation established by Haralick and Shapiro (1985), that does not require that the user set any parameter or threshold value. We identify some limitations in this evaluation function, and propose two enhanced functions that correspond more closely to visual judgment. © 1998 Elsevier Science B.V. All rights reserved.

Keywords: Color image segmentation; Segmentation evaluation; Segmentation algorithms comparison

1. Introduction

Over the past decades many low-level region segmentation algorithms have been proposed (Haralick and Shapiro, 1985; Pal and Pal, 1993). The aim of these algorithms is the domain-independent partition of the image into a set of regions that are visually distinct and uniform with respect to some property, such as gray level, texture or color. While several authors have recognized that correct segmentation can not be achieved without specific domain knowledge (e.g., Ton et al., 1991; Pavlidis and Liow, 1990), low-level segmentation is often applied as the first step in a bottom-up strategy of image analysis. In this fashion, segmentation is often evaluated only visually, or on the basis of the effectiveness of the segmentation produced in the subsequent domain-dependent step of processing (e.g. Schettini, 1993; Saber et al., 1997).

Zhang (1996) has recently published an extensive survey of existing methods for evaluating image segmentation. His analysis suggests that empirical methods are to be preferred, as there is still no general theory of image segmentation. He also points out that, in general, evaluation functions require some scaling/weighting parameters, which often have to be set on the basis of human intuition or judgment. Liu and Yang (1994) have proposed a function that does not require any user-set parameter or threshold values, for the quantitative evaluation of the performance of algorithms for the segmentation of color images. As we believe that this type of parameter-free, quantitative measure would be very

[*] Corresponding author. Tel.: +39 2 70643288; fax: +39 2 70643292, +39 2 2663030; e-mail: centaura@itim.mi.cnr.it.

[1] Electronic Annexes available. See http://www.elsevier.nl.locate/patrec.

[2] E-mail: campadelli@dsi.unimi.it.

PII: S0167-8655(98)00052-X

useful for automated applications, we have carefully tested Liu and Yang's evaluation function, and in the process identified some limitations (Borsotti, 1996; Campadelli et al., 1997). We have now gone on to design two enhanced functions that still do not require any user-set parameter or threshold value, and correspond more closely to visual judgment.

2. The evaluation function of Liu and Yang

The function proposed by Liu and Yang (1994) has been designed to incorporate, directly or indirectly, three out of the four heuristic criteria suggested by Haralick and Shapiro (1985) for evaluating the results of segmentations without having to set any threshold values for any of the subjective properties of region size, shape or homogeneity. The incorporated criteria are: (1) the regions must be uniform and homogeneous, (2) the region's interiors must be simple, without too many small holes, and (3) adjacent regions must present significantly different values for uniform characteristics. The authors, commenting on their experimental results, suggest that their function also accounts indirectly for the smoothness of the boundaries (part of the Haralick and Shapiro's fourth criterion, which includes boundary accuracy) and, as a future project, propose incorporating in an algorithm the evaluation function that guides the segmentation process itself.

Their evaluation function is empirically defined as

$$F(I) = \frac{1}{1000(N \times M)} \sqrt{R} \sum_{i=1}^{R} \frac{e_i^2}{\sqrt{A_i}}, \qquad (1)$$

where I is the segmented image, $N \times M$ the image size, and R the number of regions of the segmented image, while A_i and e_i are, respectively, the area and the average color error of the ith region; e_i is defined as the sum of the Euclidean distances between the RGB color vectors of the pixels of region i and the color vector attributed to region i in the segmented image. The smaller the value of $F(I)$, the better the segmentation result should be.

Eq. (1) is composed of three terms: the first is a normalization factor that takes into account the size of the image; the second, $\sqrt{R}$, penalizes segmentations that form too many regions; the last term, the sum, penalizes segmentations having non-homogeneous regions. Since the average color error e_i of the region is significantly higher for large regions than for small ones, e_i has been scaled by the factor $\sqrt{A_i}$.

The authors, who report a good match between function values and visual evaluation of the corresponding image segmentations, have used function F to automatically select the best segmentation with the variation of several features of their algorithm, such as the color space and the dissimilarity measure (Liu and Yang, 1994).

However, analysis of Eq. (1) shows that the presence of many regions in the segmented image is penalized only by the global measure $\sqrt{R}$. Since the average color error of small regions is often close to zero, the function tends to evaluate very noisy segmentations favorably. Consider, for example, the 200 × 200 24-bit pixel image 'Strawberry' in Fig. 1. It represents a printed fabric digitalized by a flat bed scanner, and shows regions of fairly uniform color, together with textured areas. Some segmented images, ranked according to their $F(I)$ values, are shown in Fig. 2. For details on how these segmentations were obtained, see (Borsotti, 1996). They have been ordered in ascending values of $F(I)$ from the best to the worst: Fig. 2a shows nearly all the main regions, the borders are smooth, and there are no

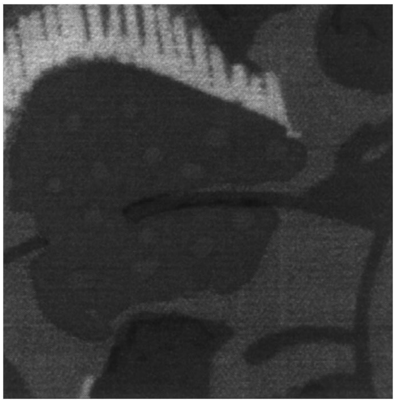

Fig. 1. Original 'strawberry' image. Available in color as an Electronic Annex (http://www.elsevier.nl/locate/patrec).

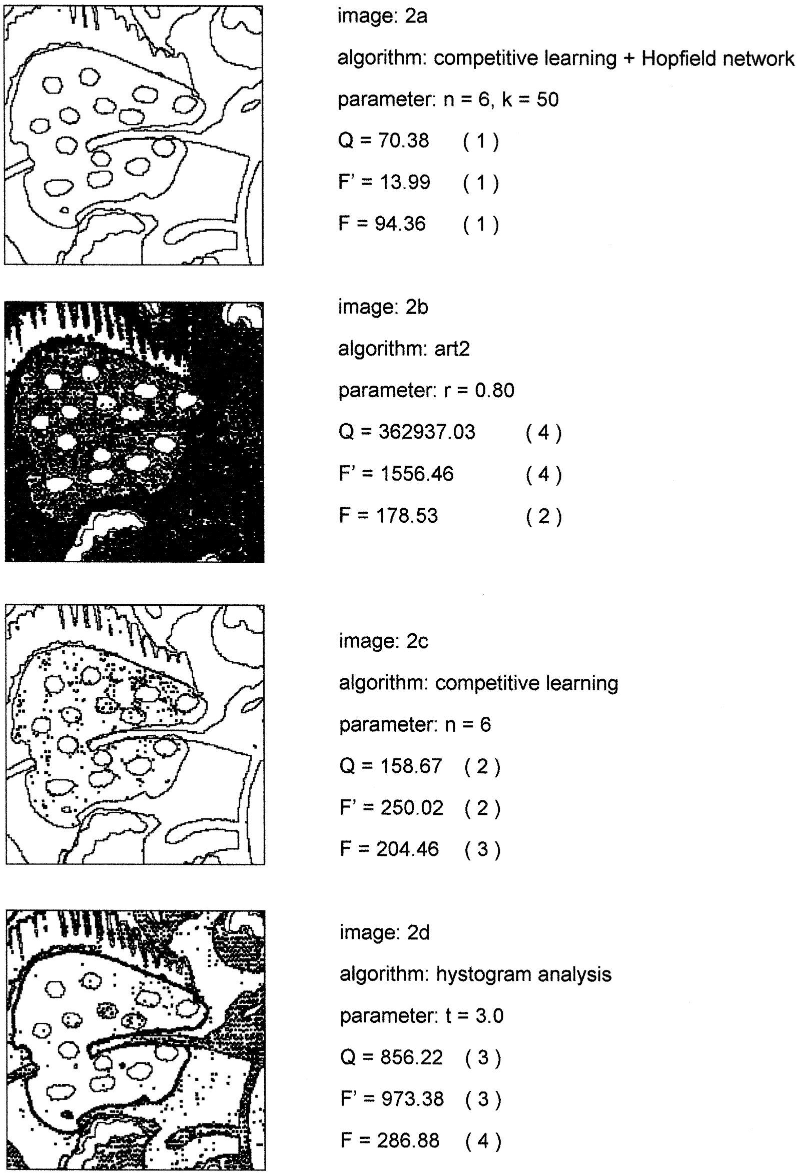

image: 2a
algorithm: competitive learning + Hopfield network
parameter: n = 6, k = 50
Q = 70.38 (1)
F' = 13.99 (1)
F = 94.36 (1)

image: 2b
algorithm: art2
parameter: r = 0.80
Q = 362937.03 (4)
F' = 1556.46 (4)
F = 178.53 (2)

image: 2c
algorithm: competitive learning
parameter: n = 6
Q = 158.67 (2)
F' = 250.02 (2)
F = 204.46 (3)

image: 2d
algorithm: hystogram analysis
parameter: t = 3.0
Q = 856.22 (3)
F' = 973.38 (3)
F = 286.88 (4)

Fig. 2. 'Strawberry' image segmentations, scores and relative ranks. Algorithms: competitive learning (Arbib and Uchiyama, 1994), Hopfield network (Campadelli et al., 1997), art2 (Carpenter and Grossberg, 1987), histogram analysis (Carlotto, 1987). Parameters: n = cluster number, k = net iterations, r = vigilance, t = histogram smoothing.

holes (small regions); Fig. 2b is actually pure noise; Fig. 2c approaches the quality of Fig. 2a, but presents some holes in the strawberry region; the image of Fig. 2d has thick borders and some very noisy regions. Considering the Haralick and Shapiro's evaluation criteria informing Liu and Yang's evaluation function, we think that, apart from any subjective ranking of Fig. 2c and Fig. 2d, Fig. 2b presents the worst segmentation and should, therefore, be ranked last instead of second.

3. The revised evaluation functions

Having ascertained that it is mainly in the evaluation of noisy segmentations that F fails to satisfy Haralick and Shapiro's criteria, our objective was to modify evaluation function F without corrupting its form and performance on not-noisy segmentations. We experimented many modifications of function F to make it penalize segmentation featuring many small regions more heavily. We began by modifying $\sqrt{R}$, the second term in Eq. (1), in order to obtain a new term with a higher value for segmented images formed by many small regions. The first, obvious attempt was to change the square root with other power functions, such as 1, 3/2, etc., but experimentation showed that this approach was not successful. We then tried another, substituting the term $\sqrt{R}$ in Eq. (1) with a new term weighting the frequency of regions' size with their respective sizes. The evaluation function thus obtained is

$$F'(I) = \frac{1}{10000(N \times M)} \sqrt{\sum_{A=1}^{\mathrm{Max}} [R(A)]^{1+1/A}} \times \sum_{i=1}^{R} \frac{e_i^2}{\sqrt{A_i}}, \tag{2}$$

where $R(A)$ is the number of regions having exactly area A, and Max the area of the largest region in the segmented image. The exponent $(1 + 1/A)$ enhances the small regions' contribution, so the sum grows as the number of small regions increases. The value of the new term is close to $\sqrt{R}$ when very few small regions are found.

The order of the segmented images of Fig. 2 according to F' was now 2a, 2d, 2b, 2c. This ranking appeared correct, but we were puzzled by the relatively large difference between the F' scores for the segmentations in Fig. 2d and Fig. 2b, which differ in reality for just a few small regions. This is due to the fact that the sum of the average color errors for the regions,

$$\sum_{i=1}^{R} \frac{e_i^2}{\sqrt{A_i}},$$

is nearly the same for both segmentations, while the new term in Eq. (2) is much larger for the segmentation with more small regions, assigning the two segmentations unrealistically different scores. The problem is that this new term is a multiplicative term *outside* the sum measuring regional color errors: the resulting function F' is not sensitive enough to small segmentation differences. Our attempts to reduce this effect by applying various decreasing functions, so that the term value decreased to zero faster, provided no solution that showed a significant improvement (Borsotti, 1996).

This observation and the fact that F', like F, has the inelegant property of reaching its minimum value (zero) on non-segmented images led us to modify the structure of F in a different way, changing the last term of function F in order to penalize both small regions and regions that have a large color error.

We experimented different solutions and found that a particularly well performing function is

$$Q(I) = \frac{1}{10000(N \times M)} \sqrt{R} \times \sum_{i=1}^{R} \left[\frac{e_i^2}{1 + \log A_i} + \left(\frac{R(A_i)}{A_i} \right)^2 \right], \tag{3}$$

where all the entities are as previously defined for F, while $R(A_i)$, as defined in F', represents the number of regions having an area equal to A_i. The body of the sum is composed of two terms: the first is high only for non-homogeneous regions (typically, large ones), while the second term is high only for regions whose area A_i is equal to the area of many other regions in the segmented image (typically, small ones). In designing this new term we took into account the fact that we may expect that the number of regions of area A_i in given an image *will be* small if area A_i has a high value; and in this case $R(A_i)/A_i$ contributes little to the sum. On the other hand, the number of regions of area A_i *may* be large if the area A_i has a low value; in this case $R(A_i)/A_i$ contributes strongly to the sum. Heuristically we can say that $R(A_i)$ is almost always 1 for large regions, and can be much larger than 1 for small regions. In any case, the denominator A_i drastically forces the term $R(A_i)/A_i$ to near zero for large regions, and lets it grow for small regions.

Two further modifications of F were made to obtain Q. The first term in the sum also differs from its corresponding term in F: $\sqrt{A_i}$ has been replaced with $(1+\log A_i)$ to obtain a stronger penalization of non-homogeneous regions. Finally, the arbitrary normalization term has been scaled by 10 to obtain a range of values similar to those of F and F'.

The evaluation given by Q is shown in Fig. 2: the order is the same as that obtained by F', but the scores better reflect the visual evaluation of segmentation 'quality'. Fig. 2b, in particular, has a very high score, meaning that it should not be considered for further processing.

4. Experimental results

We have compared the ranking performance of F, F' and Q on ten test images processed by six clustering methods – multithresholding by a histogram analysis algorithm (Carlotto, 1987), a competitive learning clustering algorithm (Arbib and Uchiyama, 1994), two Adaptive Resonance Theory (ART) based algorithms (Baraldi and Parmiggiani, 1995; Carpenter and Grossberg, 1987), a Reactive Tabu Search (RTS) based algorithm (Al-Sultan, 1995) and the widely used Isodata algorithm (Ball and Hall, 1967) –, and three additional methods which use spatial information to enhance the results of the clustering – two Hopfield neural networks (Campadelli et al., 1997; Yu and Tsai, 1991), and a constraint satisfaction neural network (Lin et al., 1992). Readers desiring more detailed information about the algorithms implemented are referred to (Borsotti, 1996).

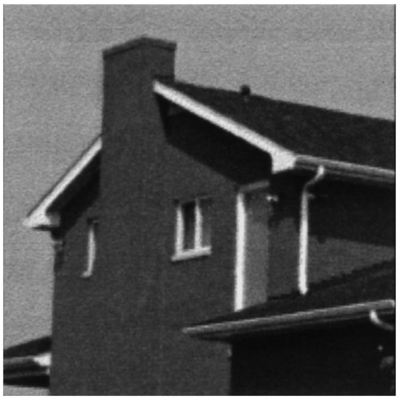

Fig. 3. Original 'house' image. Available in color as an Electronic Annex (http://www.elsevier.nl/locate/patrec).

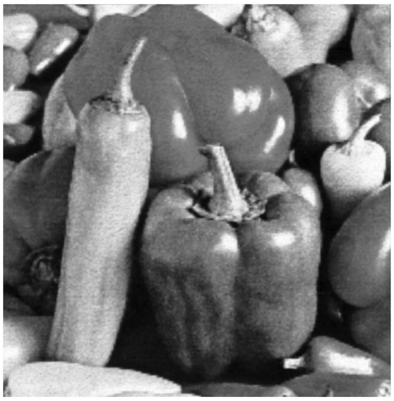

Fig. 4. Original 'peppers' image. Available in color as an Electronic Annex (http://www.elsevier.nl/locate/patrec).

For each test image we applied each segmentation algorithm in turn, varying the input parameters (such as the number of clusters) and then compared the ranking obtained by applying F, F' and Q with the results of visual evaluation (in agreement with Pal and Pal (1993) who write in their review of image segmentation techniques ''a human being is the best judge to evaluate the output of any segmentation algorithm'').

The F evaluation function generally ranks 'good' segmentations results correctly, but overrates 'bad' – typically, noisy – segmentations. F' produces the same ranking as Q, and this agrees with the subjective visual order. However, the variations in Q values match more closely than those in F' the corresponding variations in visual judgment.

Typical results for the 'house' and 'peppers' images (Figs. 3 and 4) are given in Figs. 5 and 6, where the values of F, F' and Q and the relative ranking positions are shown alongside the images. Fig. 5a–d shows four segmentations of the 'house' image obtained by the competitive learning clustering algorithm of Arbib and Uchiyama (1994) with 5, 6, 7 and 8 color clusters, respectively. Fig. 6a–d shows four segmentations of the 'peppers' image obtained with the Simplified Adaptive Resonance Theory algorithm (Baraldi and Parmiggiani, 1995). According to this algorithm a single vigilance parameter, $r \in [0,1]$,

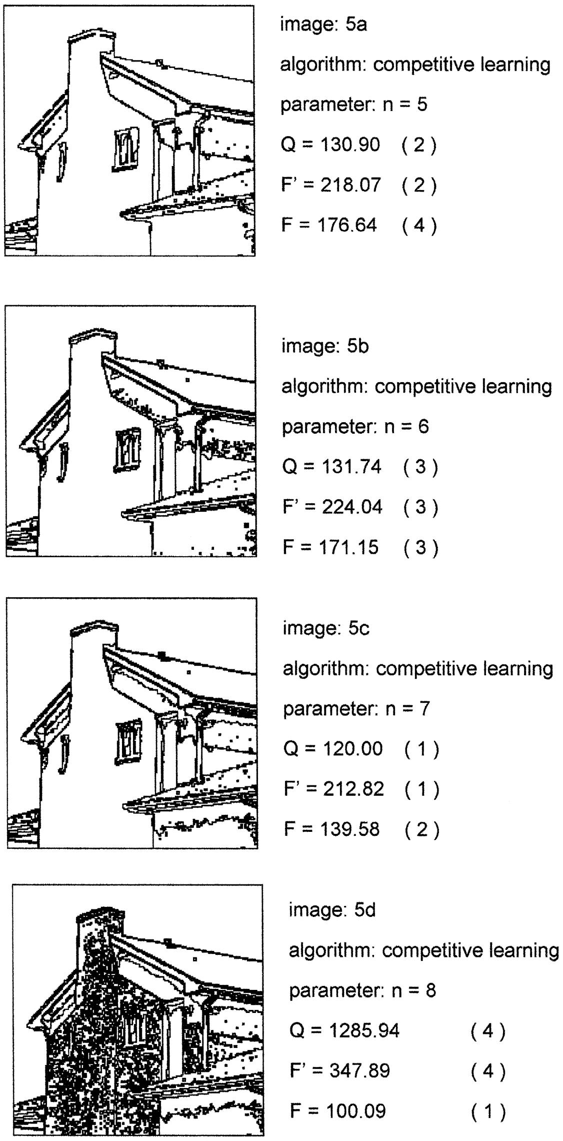

image: 5a
algorithm: competitive learning
parameter: n = 5
Q = 130.90 (2)
F’ = 218.07 (2)
F = 176.64 (4)

image: 5b
algorithm: competitive learning
parameter: n = 6
Q = 131.74 (3)
F’ = 224.04 (3)
F = 171.15 (3)

image: 5c
algorithm: competitive learning
parameter: n = 7
Q = 120.00 (1)
F’ = 212.82 (1)
F = 139.58 (2)

image: 5d
algorithm: competitive learning
parameter: n = 8
Q = 1285.94 (4)
F’ = 347.89 (4)
F = 100.09 (1)

Fig. 5. ‘House’ image segmentations, scores and relative ranks. Algorithms: competitive learning (Arbib and Uchiyama, 1994). Parameter: $n =$ cluster number.

regulates the sensitivity to color differences: a higher value corresponds to a finer classification of the color images in clusters, although the exact correspondence between the vigilance parameter and the resulting number of clusters can not be predicted. The first two segmentations (Fig. 6a and Fig. 6b) were obtained with the vigilance parameters set at 0.75 and 0.80, respectively; the last two (Fig. 6c and Fig. 6d) with the vigilance parameter at 0.85. The segmentation in Fig. 6d was then post-processed by a Hopfield neural network to enhance the results of the clustering with spatial information (Campadelli et al., 1997). F' and the Q ranked the segmentations of these images in the same way, and this was also the

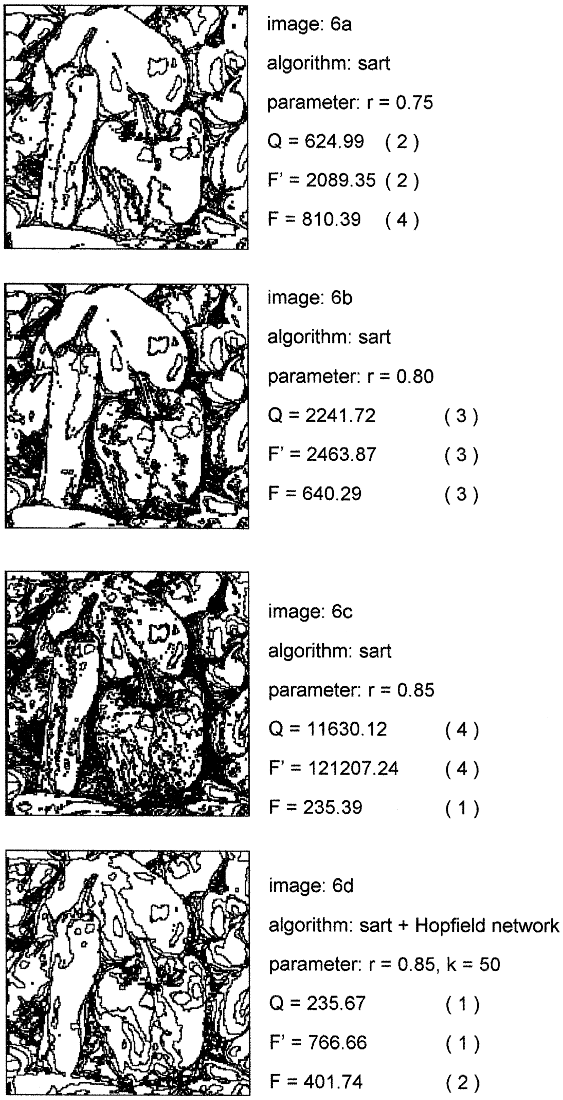

image: 6a
algorithm: sart
parameter: r = 0.75
Q = 624.99 (2)
F’ = 2089.35 (2)
F = 810.39 (4)

image: 6b
algorithm: sart
parameter: r = 0.80
Q = 2241.72 (3)
F’ = 2463.87 (3)
F = 640.29 (3)

image: 6c
algorithm: sart
parameter: r = 0.85
Q = 11630.12 (4)
F’ = 121207.24 (4)
F = 235.39 (1)

image: 6d
algorithm: sart + Hopfield network
parameter: r = 0.85, k = 50
Q = 235.67 (1)
F’ = 766.66 (1)
F = 401.74 (2)

Fig. 6. ‘Peppers’ image segmentations, scores and relative ranks. Algorithms: Hopfield network (Campadelli et al., 1997), sart (Baraldi and Parmiggiani, 1995). Parameters: $k =$ net iterations, $r =$ vigilance.

case of the more than 500 segmented color images we tested. For this reason ranking alone is not a criterion in choosing between the two functions. However, as said above, the variations in the Q score correspond more closely than those of F' to the visual evaluation of the different segmentations. For example, the ratio between the F' scores for Fig. 2a and Fig. 2c is about twenty, and only about six for Fig. 2b and Fig. 2c, while the corresponding Q ratios are, respectively, two and over 2200, an estimation closer to subjective judgment. Fig. 5a, Fig. 5b and Fig. 5c are visually similar, and the values obtained with both F' and Q reflect this. However, for the image of Fig. 5d, F' registers a score only 1.5-times the score of the images of Fig. 5a, Fig. 5b and Fig. 5c, while the Q value for Fig. 5d is about 10-times the Q values for Fig. 5a, Fig. 5b and Fig. 5c, again showing a closer correspondence to visual judgment.

The low-level segmentations of the ‘pepper’ image are much harder to evaluate subjectively because of the intrinsic difficulty of judging the low-level segmentations of an image that presents so many shades of color, shadows and highlights: in this case, while the ranking appears correct, it is practically impossible to correlate between the numerical value of the scores with the ‘quality’ of the segmentation.

5. Conclusions

There is a growing demand in image analysis for an automation of the evaluation of low-level segmentations that does not require that the user set any parameter or threshold values. Unfortunately, few evaluation schemes of this kind have been developed (Zhang, 1996). The function F, proposed by Liu and Yang (1994), represents an elegant attempt to evaluate color segmentation results quantitatively and objectively. In this note, after a critical analysis of the strengths and limitations of evaluation function F, we have proposed two enhanced functions F' and Q, which retain all the merits of the F function while eliminating the drawbacks. The results of our experimentation suggest that both F' and Q can successfully pick out the ‘best’ in a set of possible segmentations, but Q should be preferred to F' as a guide for tuning segmentation algorithms.

References

Al-Sultan, K.S., 1995. A tabu search approach to the clustering problem. Pattern Recognition 28 (9), 1443–1451.

Arbib, M.A., Uchiyama, T., 1994. Color image segmentation using competitive learning. IEEE Trans. on Pattern Analysis and Machine Intelligence 16 (12), 1197–1206.

Ball, G.H., Hall, D.J., 1967. A clustering technique for summarizing multivariate data. Behav. Sci. 12, 153–155.

Baraldi, A., Parmiggiani, F., 1995. A neural network for unsupervised categorization of multivalued input patterns: an application to satellite image clustering. IEEE Trans. on Geoscience and Remote Sensing 33 (2), 305–316.

Borsotti, M., 1996. Segmentazione di immagini a colori mediante clustering. Università degli Studi di Milano, Tesi di Laurea in Scienze dell’Informazione, A.A. 1995–1996.

Carlotto, M.J., 1987. Histogram analysis using a scale-space approach. IEEE Trans. on Pattern Analysis and Machine Intelligence 9, 121–129.

Carpenter, G.A., Grossberg, S., 1987. ART2: self-organization of stable category recognition codes for analog input patterns. Applied Optics 26, 4919–4930.

Campadelli, P., Medici, D., Schettini, R., 1997. Color image segmentation using Hopfield networks. Image and Vision Computing 15, 161–166.

Haralick, R.H., Shapiro, L.G., 1985. Image segmentation techniques. Computer Vision Graphics Image Processing 29, 100–132.

Lin, W.C., Tsao, E.C.K., Chen, C.T., 1992. Constraint satisfaction neural networks for image segmentation. Pattern Recognition 25 (7), 679–693.

Liu, J., Yang, Y.-H., 1994. Multiresolution color image segmentation. IEEE Trans. on Pattern Analysis and Machine Intelligence 16 (7), 689–700.

Pal, N.R., Pal, S.K., 1993. A review on image segmentation techniques. Pattern Recognition 26 (9), 1294–1294.

Pavlidis, T., Liow, Y.-T., 1990. Integrating region growing and edge detection. IEEE Trans. on Pattern Analysis and Machine Intelligence 12 (3), 225–233.

Saber, E., Murat Tekalp, A., Bozdagi, G., 1997. Fusion of color and egde information for improved segmentation and edge linking. Image and Vision Computing 15, 769–780.

Schettini, R., 1993. A segmentation algorithm for color images. Pattern Recognition Letters 14 (6), 499–506.

Ton, J., Stricklen, J., Lain, A.K., 1991. Knowledge-based segmentation of landsat images. IEEE Trans. on Geoscience and Remote Sensing 29 (2), 222–232.

Yu, S.S., Tsai, W.H., 1991. Relaxation by the Hopfield neural network. Pattern Recognition 25 (2), 197–209.

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