Unsupervised Colour Image Segmentation by Low-Level Perceptual Grouping

Abstract   This paper proposes a new unsupervised approach for colour image segmentation. A hierarchy of image partitions is created on the basis of a function that merges spatially connected regions according to primary perceptual criteria. Likewise, a global function that measures the goodness of each defined partition is used to choose the best low-level perceptual grouping in the hierarchy. Contributions also include a comparative study with five unsupervised colour image segmentation techniques. These techniques have been frequently used as a reference in other comparisons. The results obtained by each method have been systematically evaluated using four well-known unsupervised measures for judging the segmentation quality. Our methodology has globally shown the best performance, obtaining better results in three out of four of these segmentation quality measures. Experiments will also show that our proposal finds low-level perceptual solutions that are highly correlated with the ones provided by humans.

1 Introduction

Image segmentation refers to the process of partitioning an image into several non-intersecting regions that hopefully correspond to structural units in the scene or any object of interest. Each region will be made of connected pixels and will be homogeneous according to certain criteria (intensity, texture, motion, etc.) and the union of all the regions forms the whole image [1].

In the last few decades, many researchers have focused their work on algorithms and techniques that look for the main regions that compose an image [2–4]. This means that the state-of-the-art on image segmentation involves large amounts of methodologies and also many taxonomies on this topic [5–7]. From these taxonomies, one of the early and extensively used methodologies in the literature is the hierarchical one since identifying structures on an image is inherently a multiscale problem. Thus, multiscale approaches from more than

a decade ago [8] are still present in nowadays works [9], contributing to make these approaches more effective. A hierarchical methodology is mostly designed as an optimisation problem to reach general sub-optimal values for an objective function that measures the "quality" of an image partition. Moreover, hierarchical approaches are commonly combined with a similarity criterion between regions that uses all the information extracted from the regions in order to decide if they should be merged or not in a region growing scheme [10, 11]. Two main reasons lead us to choose a hierarchical strategy in our approach:

– General-purpose image segmentation techniques deal with different sources and/or scenes. This variety of tasks can be faced up in a very intuitive way by means of hierarchical methods. Likewise, these data-adaptive structures provide a description of a scene in terms of regions at several resolution levels [12].
– Hierarchical structures produce a multiscale representation which allows to design efficient and easy implementations that result in robust algorithms.
– Region grouping processes are more appropriate than clustering or thresholding approaches since they simultaneously take into account both colour information and its distribution in the spatial domain [13].

There have also been many attempts to achieve the optimal image segmentation result according to certain perceptual premises [14, 15]. Many discussions on *what is* or *how to get* natural segmentation results have been published [16] and a significant effort has been devoted to developing complex scene image segmentation [17, 18], perceptual grouping of image contents [19, 20] or perceptually based colour texture analysis [21]. A lot of work in this direction has also been based on perceptual colour spaces [22]. In our case, being aware that both approaches are not mutually exclusive, we have rather focused on the idea that perceptual organisation of regions in an image is the key point in order to *mimic* object recognition process of humans [23, 24]. Therefore, it is important to measure how the pixels are distributed within each region and to take into account the relationship among

the regions in the whole image. This is a very challenging problem that has motivated our work on extracting low-level image features that can be correlated with high-level semantics.

Another important issue in this approach comes from the difficulty of making any a priori assumption in the frame of a general approach for image segmentation. Unsupervised methods, unlike the supervised ones, avoid any kind of prior knowledge. This characteristic is indispensable to perform general-purpose applications, such as graphics editing programs, off-line image analysis or simply as a pre-processing step for further high-level tasks. The ability to work without making a priori assumptions allows unsupervised methods to operate over a wide range of conditions and with many different types of images.

Evaluating the quality or the impact of the results is a critical point in any scientific work, being even more important in image segmentation due to the subjective evaluations that this task often involves. Thus, the approach on this point should be carefully designed. In the often-cited work of Zhang [25], the different methodologies to evaluate segmentation algorithms are broadly divided into two categories: *analytical methods* and *empirical methods*. Due to the difficulty to compare algorithms solely by means of analytical studies, analytical methods have not received much attention in the literature [26]. There also exists a considerable amount of empirical methods which are, in turn, classified into two types: *discrepancy methods* and *goodness methods*. An evaluation criterion belongs to the empirical discrepancy methods or to the empirical goodness methods depending on whether the gold-standard or ground-truth reference is available or not respectively. In the framework of our current projects, expecting to have a ground-truth reference of the *ideal* segmentation result is not always possible and, therefore, we discard the empirical discrepancy as the evaluation criterion. Consequently, we will focus on the empirical goodness measures, that are often called *unsupervised evaluation methods*, for estimating the quality of the results.

Recent reviews on unsupervised evaluation methods [27, 28] compare the performance of these measures by means of different experiments and suggest which ones work better on each scenario. Moreover, authors in [29] support the idea of taking a collection of evaluation criteria in order to avoid the bias of the individual measures. Thus, we have taken advantage of the conclusions of these works in order to select and apply those criteria that perform the best to measure the performance of the algorithms.

The goal in this work is to develop a general-purpose colour image segmentation algorithm. In order to obtain this objective, a hierarchical structure of partitions in the image domain is proposed [30, 13]. This hierarchy is developed in a fully unsupervised way based on two novel criteria for grouping regions and choosing the most suitable image partition in the hierarchy. These criteria use primary perceptual principles as achieving maximum contrast among regions while preserving intra-region colour homogeneity and edge information. The segmentation process starts from an over-segmented representation of the image, which constitutes the first level of the hierarchy. From this partition, adjacent regions are progressively merged according to a grouping criterion that selects the two most similar regions. The merging procedure is repeated to produce successively more levels until only one region covers the whole image. Finally, another criterion is used to select the best level from the hierarchy. The selected partition is understood as a preliminary step towards the semantic grouping that a human would made starting from our resulting low-level perceptual grouping.

Another main contribution of this work is a systematic comparison of the most successful colour image segmentation algorithms. These algorithms have been widely used in the literature due to their reasonable performance. Besides this, their source code is freely available. A systematic comparison among this wide range of algorithms expands the state-of-the-art on this field providing objective information about their performance.

The remainder of this paper is organised as follows: In Section 2, the proposed image segmentation algorithm based on the optimisation of a criterion function is described. Experimental results, comparisons and analysis of the results are presented in Section 3. Section 4 concludes the paper and discusses future work.

2 The proposed algorithm

Psychological approaches based on Gestalt laws agree on the importance of the region homogeneity measurement and the edge extraction procedures in the human visual system [24]. The effectiveness of combining these two aspects has been already demonstrated in other algorithms for perceptual image segmentation [15]. The proposed grouping criterion is focused on the spatial features of the perception of the different regions that form the image rather than on the representation spaces. Therefore, it can be applied in any representation space, from grey to colour images or even to multi/hyperspectral images, which is an important quality in the framework of our current research projects. Thus, we will concentrate our effort on those measures related to grouping and on choosing a partition with a suitable spatial distribution.

The image segmentation method that we propose here employs a hierarchical methodology in an agglomerative way. Thus, starting from a highly over-segmented representation of the image, the proposed algorithm will group together those spatially connected regions that are similar enough according to the criterion function exposed in Section 2.1. This iterative process based on primary perceptual criteria will create a hierarchy of partitions. Another function will measure the goodness for each partition in the hierarchy. This second function will be explained in Section 2.2.

At first sight, the algorithm could only be based on the optimisation of the functional that performs the clustering assessment. However, each functional measures distinct features and, although the second functional selects the best par-

tition, the first one decides how the hierarchical structure is scanned through [10].

Gaussian models have been successfully used in many works on segmentation using natural images [31–34]. In our approach we also assume a Gaussian distribution of the region pixel values.

2.1 Defining a dissimilarity measure between regions

Let us suppose an initial coarse representation of the image where each homogeneous group of connected pixels is treated as a single region. From this initial partition, pairs of regions are successively merged until all the regions have been merged into just one, creating a hierarchical process that follows a single-link strategy. This means that each iteration takes into account the dissimilarity (D) between every pair of spatially connected regions, being the two most similar regions according with this criterion forced to merge. Therefore, our strategy implements a deterministic sequence of merging operations, where these merging operations are irreversible.

Let us assume a set of $\mathcal{B}$ bands that form a multiband input image f, being $f(x)$ a $\mathcal{B}$-dimensional vector associated with each $\mathcal{B}$-dimensional pixel x. In our proposal, measure D is defined as follows:

$$D_{RR'} = d_{RR'}(1 + \delta_{RR'}) \tag{1}$$

where

$$d_{RR'} = (\mu_R - \mu_{R'})^T \left(Cov_R^{-1} + Cov_{R'}^{-1}\right)(\mu_R - \mu_{R'}) \tag{2}$$

and

$$\delta_{RR'} = \frac{\sum_{x \in B_{RR'}} max_{i=1...\mathcal{B}} \left(|\nabla f_i(x)|^2\right)}{card\left(B_{RR'}\right)} \tag{3}$$

Mean values of the pixels of regions R and R' are represented by vectors μ_R, $\mu_{R'}$ whereas the covariance matrices are represented by Cov_R, $Cov_{R'}$, The dimensionality of the mean vectors and covariance matrices depends on $\mathcal{B}$. In the $\delta_{RR'}$ term, $B_{RR'}$ is the set of pixels that belong to the boundary between regions R and R'. Function $card(\cdot)$ returns the cardinality of a set. Rather than using a more sophisticated method, a maximum value of the gradient magnitude found in each band is used. Thus, $|\nabla f_i(x)|$ is the magnitude of the gradient at point x for the band i and function $max(\cdot)$ returns the maximum of the values in brackets.

Under the assumption of considering normally distributed pixels in each region, the value of $d_{RR'}$ is a Mahalanobis distance between two distributions, defined in the pixel domain. This term accounts for the similarity in the distribution of pixel values in the two regions that are considered to be merged. The term $\delta_{RR'}$ averages the square of the gradient magnitude values along the boundary between these two connected regions. Thus, this term accounts for the strength of the discontinuity between the distributions of pixel values across the regions under consideration, including a spatial measure of discontinuity in the dissimilarity function.

Dissimilarities in Eq. (2) and Eq. (3) are not complementary but parallel events. If both events happen, dissimilarity $D_{RR'}$ in Eq. (1) is reinforced. $D_{RR'}$ takes into account $d_{RR'}$, the dissimilarity between region distributions itself, and $d_{RR'} \cdot \delta_{RR'}$, the product of the dissimilarities between region distributions and the edge strength at the border of both regions. Thus, if the difference between the pixel distributions in R and R' is high, $D_{RR'}$ value still gives more importance to the edge between them. If this difference is low, the edge could be due to the presence of texture and its importance is reduced.

2.2 Deciding the number of regions

A criterion for selecting the most suitable partition in the hierarchy is described in this section. It is completely independent of the measure defined in Equation (1) for grouping regions and regardless of the way in which the hierarchical tree structure is constructed. Thus, while the algorithm progressively merges pairs of neighbouring regions, a non-parametric estimation of the goodness of the data partition is performed. The resulting partition is selected without any a priori information about the final number of regions or the shape of the resulting ones.

The algorithm has been formulated as the maximisation of a criterion function F. This function attempts to quantify the basic perceptual principle of maximising the contrast in the image domain while preserving intra-region colour homogeneity. That is, given an image partition, we measure how well the pixels fit their corresponding regions and not the neighbouring ones. Therefore, the algorithm will select the partition where the value of the F function would be maximum:

$$F = S_i \cdot S_e \tag{4}$$

$$S_i = \frac{1}{N} \sum_R \sum_{x \in R} S(x, R) = \frac{1}{N} \sum_R \sum_{x \in R} e^{\frac{-|f(x) - \mu_R|^2}{2\sigma^2}} \tag{5}$$

$$S_e = \frac{1}{N} \sum_R \sum_{x \in R} \bar{S}(x, \mathcal{N}(R)) =$$

$$= \frac{1}{N} \sum_R \sum_{x \in R} \prod_{R'}^{\mathcal{N}(R)} \bar{S}(x, R') =$$

$$= \frac{1}{N} \sum_R \sum_{x \in R} \prod_{R'}^{\mathcal{N}(R)} (1 - S(x, R')) =$$

$$= \frac{1}{N} \sum_R \sum_{x \in R} \prod_{R'}^{\mathcal{N}(R)} \left(1 - e^{\frac{-|f(x) - \mu_{R'}|^2}{2\sigma^2}}\right) \tag{6}$$

For each partition in the hierarchy, S_i is an inner measure and S_e an external measure. In these equations, N represents the total number of pixels in the image. R, R' are regions and $\mathcal{N}(R)$ is the set of neighbouring regions of region R. The

value of the pixel x is represented by $f(x)$ whereas μ_R is a vector containing the average value of the pixels in the region R. $S(x, R)$ is a measure of similarity between pixel x and region R.

Both Equation (5) and Equation (6) assume a Gaussian distribution characterised by a mean and an expected variance. $\bar{S}(x, \mathcal{N}(R)) = \prod_{R'}^{\mathcal{N}(R)}(1 - S(x, R'))$ is the complement of the measure that a pixel belongs to the neighbouring regions $\bar{S}(x, \mathcal{N}(R))$, which is defined as the measure that a pixel does not belong to any of the neighbouring regions.

Equations (5) and (6) are inspired by probability calculus, although we are aware that we are not strictly dealing with probabilities but non-normalised probabilities. S_i can be considered as the average measure of the non-normalised probability that a pixel in the image belongs to the region that it has been assigned to. On the contrary, S_e is the average measure of the non-normalised probability that a pixel does not belong to its neighbouring regions. The criterion function F therefore takes into account the fact that the pixels belong to the assigned regions in the partition and, simultaneously, that the pixels do not belong to neighbouring regions in the image domain. Hence, this estimates how "well" the pixels are grouped and whether these groups are internally consistent and, at the same time, different enough from spatially nearby regions.

It is worth saying that not only the optimal partition could be selected by means of function F but other suboptimal partitions could be taken into account by means of selecting other local maxima of the functional or selecting the partition with less regions after a period of merging operations where the functional F has remained stable. This could also end up in satisfactory partitions depending on the application and, in fact, according to our experience, this usually happens.

Parameter σ is a constant that acts as a bound on the variability of the regions and represents a smoothing constraint of the expected segmentation. The lower σ is, the larger amount of regions will have the chosen partition using Equation (4). It is important to note that this threshold is related to the variance allowed in the density estimation of the regions and it is not used as a grouping criterion. In (1), the variability of the regions is taken into account, however, in (5) and (6), σ is a constant that imposes an equal limit spread in the colour space of the regions. Otherwise, if σ is not constant and adapts to each region variability during the evaluation process, the function F will always grow and no clear maximum will be reached, becoming a degenerated problem.

The proposed approach is not significantly dependent on the choice of this σ value since the criterion function F has demonstrated a quite robust behaviour if σ is slightly changed. This is due to the fact that the sequence of merging operations does not change because it is based on the function (1) and the process described in Section 2.1. Therefore, merging two regions in a correct way still increases the function F whereas wrongly merging operations are still penalised, that is, local maxima are preserved in any case independently of the variations of σ. A fact that supports this behaviour is that it was not difficult to select a common σ value for all the images of the Berkeley segmentation database. In the experimental part of this work, σ has been fixed to the same value for all the experiments in order to not take advantage of tuning up this parameter in each experiment.

Algorithm 1 shows the pseudo-algorithm that summarises the proposed methodology.

Algorithm 1 Pseudo-Algorithm of the segmentation process.

Input: partition $\mathcal{P}$
Output: partition $max\mathcal{P}$
1 : $max\mathcal{P} = \emptyset$
2 : $max\mathcal{F} = 0$
3 : Compute $NRegs = card(P)$
4 : while $NRegs > 1$:
5 : Compute F for partition $\mathcal{P}$ (Eq. (4))
6 : if $F > max\mathcal{F}$:
7 : $max\mathcal{P} = P$
8 : $max\mathcal{F} = F$
9 : for any pair of neighbouring regions R, R':
10 : Compute $D_{RR'}$ (Eq. (1))
11 : Merge R, R' whose $D_{RR'}$ is minimum to obtain new $\mathcal{P}'$
12 : $\mathcal{P} = \mathcal{P}'$
13 : $NRegs = NRegs - 1$
14 : Return partition $max\mathcal{P}$

---

3 Experiments and results

A varied number of experiments to assess the performance of the algorithm have been carried out. Image segmentation results will be evaluated *i)* using real images from a well-known database, *ii)* comparing the performance against other renowned image segmentation algorithms [35–39] and *iii)* by means of different unsupervised evaluation criteria. These criteria were recently employed in several comparisons, obtaining an excellent performance in almost all the tests [27, 28]. The value of parameter σ in Equations (5) and (6) was fixed to 1.65 for all the experiments.

3.1 Preliminary notes and some examples

The method here presented works with RGB images. It develops a coarse-to-fine segmentation strategy based on a process that progressively agglomerates the initial regions. As starting point, an over-segmented representation of the input image is somehow needed. This representation must be over-segmented enough to ensure that all the details have been captured. Note that any wrong merging operation in the initialisation stage cannot be recovered later. In our case, this initial segmentation is performed by using a previous work [40]. However, any initial partition containing all the important details of the image would be valid as well.

Figure 1 shows a segmentation result for the *woman* image with its initial over-segmented representation. Although

this initialisation is quite poor, the final result has not been affected in a significant way. In addition, Figure 2 shows the graphical representation of the F function and its S_i, S_e components for the same image. It is worth noticing that, from our experience, any result for the F criterion where $S_i \cdot S_e \leq 0.2$ should be generally discarded. Low values for the function F are very uncommon, although they could happen on textured images where the colour palette among regions is very similar. These very low values would actually indicate a weak homogeneity relationship among the pixels of each region which, at the same time, would not be too different from the ones of the neighbouring regions. Functional F reaches its maximum value when the number of regions is equal to 9, fourth resulting image in Fig. 1. The segmentation result for 23 regions is also offered in Fig. 1 (third image) as a matter of comparing two segmentation results that obtain very similar values for functional F.


Fig. 1 Segmentation result for the *woman* image. From left to right, original *woman* image, initialisation and segmentation results at 23 regions and 9 regions (where function F reaches its maximum value).
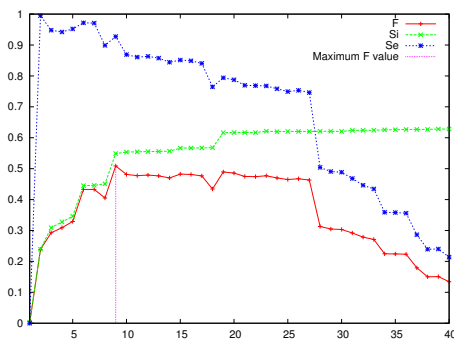
Fig. 2 Goodness estimation for the colour image of *woman*. y-axis represents the numerical result of applying the criterion function F to each partition whereas x-axis shows the number of regions of the partition. Function F reaches the maximum value at 9 regions.

Figure 3 shows the segmentation results of four classical colour images, namely, *toys*, *peppers* and *tree*. As we can see, the results in all the images are consistent with a perceptual interpretation: contours are well defined and all important regions have been detected.
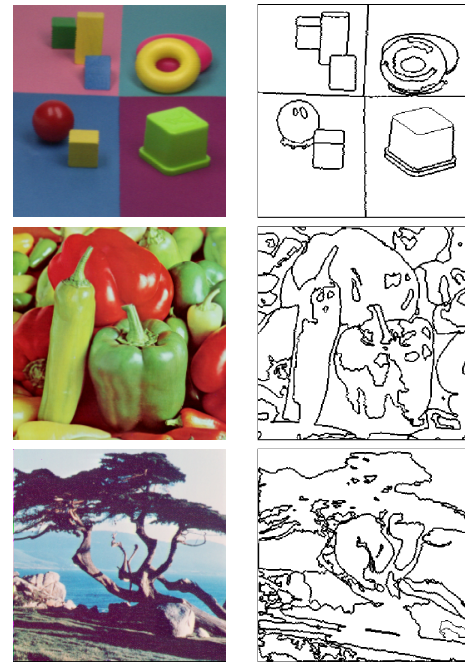
Fig. 3 Examples of results on classical images: from top to bottom, toys, peppers and tree.

It is important to notice that hereafter we will refer to our proposal as **PSEG** algorithm, as the acronym of *Partition*-based *SEGmentation*.

3.2 Overview of other segmentation algorithms and evaluation measures

In addition to the proposed algorithm, five unsupervised colour segmentation algorithms are used for comparison. All of them have no particular application field, being their results often used as a reference to beat in other comparatives.

1. Firstly, **MS** is an effective algorithm that can be used to obtain the dominant colours of an image using the CIE $L*u*v*$ colour space. It was proposed by Comaniciu and Meer in [35] and it is based on the *mean shift* algorithm applied in the spatial domain.
2. In [36] Felzenszwalb and Huttenlocher presented an algorithm (**FH**) that adaptively adjusts the segmentation criterion based on the degree of variability in neighbouring regions of the image. Simultaneously, a graph-based approach guides the segmentation process. The algorithm starts a coarse-to-fine iterative process until the stage where the resulting partition is neither too coarse nor too fine.
3. Statistical Region Merging (**SRM**) algorithm was proposed by Nock and Nielsen in [37] based on the idea of using perceptual grouping and region merging for image segmentation. Our proposal is based on a similar approach although using totally different merging and stopping criteria.
4. **JSEG** algorithm proposed by Deng and Manjunath in [38] provides colour-texture homogeneous regions which

are useful for salient region detection. The algorithm calculates distances between regions on the CIE $L*u*v*$ colour space. It has been widely used in natural image segmentation so far.

5. Authors in [39] have recently proposed an unsupervised colour image segmentation algorithm (**GSEG**) that is primarily based on colour-edge detection, dynamic region growth and in a multi-resolution region merging procedure, exploiting the information obtained from the CIE $L*a*b*$ colour space. This algorithm was tested on the same database of images used in this work.

It is important to point out that the parameters of the algorithms have been set to the default values as authors provide and/or suggest, therefore, no parameters have been tuned up.

Many proposals have been published about measuring the quality of segmentation results [26,27,41]. This is not an easy task since evaluating image segmentation results must be considered as a top-down problem that very often introduces an important element of subjectivity in the evaluation process. In our case, we have reduced the influence of this inconvenience by means of using a varied range of unsupervised methods for evaluating the segmentation quality of the results. Although, there exists a general agreement in the literature about the need of these quantitative measures, there is currently no consensus on which one should be used. Therefore, many alternatives for the estimation of segmentation quality could be taken but, at the same time, there is no perfect way to perform this evaluation due to the subjectivity inherent in image segmentation.

The quality measures used in this work have been used in several reviews [27,28]. In [28], authors carried out several experiments comparing 8 evaluation measures. These measures implement different criteria in order to quantify the goodness of the partitions obtained by the segmentation algorithms. On the basis of this work, we will use 4 of these measures. With the same nomenclature as in [28], these measures are E, E_{CW}, Z_{eb} and F_{RC}, which can be described as follows[1]:

E [42] : This evaluation function is based on information theory and the minimum description length principle. It uses *region entropy* as its measure of intra-region uniformity. It also uses *layout entropy* to penalize over-segmentation when the region entropy becomes small. CRITERION: the *lower* the value, the *better* the result.

E_{CW} [43] : This measure is a composite evaluation method for colour images (sum of two measures) that is based on the use of visible colour difference. It uses an intra-region error to evaluate the degree of under-segmentation, and uses an inter-region region error to evaluate the degree of over-segmentation. This measure is defined over the CIE $L*a*b*$ colour space. CRITERION: the *lower* the value, the *better* the result.

Z_{eb} [44] : This measure takes into account the internal and external contrast of the regions measured in the neighbourhood of each pixel. The internal contrast is normalised by the number of pixels of each region whereas the external contrast is normalised by the number of pixels in the region perimeter. CRITERION: the *higher* the value, the *better* the result.

F_{RC} [45] : This evaluation criterion takes into account the global intra-region homogeneity and the global inter-region disparity. The intra-region disparity is the standard deviation of the pixel values of each region. The inter-region disparity is the average of a distance between the current region and all its neighbouring regions. This distance is related to the average of the grey-level of each region. CRITERION: the *higher* the value, the *better* the result.

There are other measures in [28] that have not been used in our experiments because they focus on video sequences, are based on shape regularity, or provided poor performance in other studies [27,42].

In their review, Zhang et al. [28] also compared the measures on different environments, confirming that each measure generally works better in a different context. F_{RC} and Z_{eb} demonstrated to be the best ones in most of the cases. Nevertheless in our case, it is especially interesting that the E and F_{RC} measures were proved as the best ones by far when the machine segmentations were compared against the segmentation results specified by humans. This could be taken as the closest example to the perceptual case.

3.3 Comparing the image segmentation algorithms

Authors in [23] define perceptual information and use this concept to conclude that using just low-level image features is not possible to achieve segmentation performance comparable to human segmentation. However, there is not doubt about the suitability of human segmentations as a reference result for the image segmentation algorithms. Human-based segmentations results will surely have a more meaningful contents at least semantically.

In this section the colour images of natural scenes from the Berkeley segmentation database (BSD) will be used [46]. This database offers a set of test images and, for each image, several segmentation results labelled by humans are also provided. Manually segmented images have been often used as a perceptual reference for comparison purposes [15,23], considering the more similar to this reference, the better the segmentation result. Although BSD provides a set of 300 images, we have used their list of a hundred images ranked according to the relative complexity of each image (validation set). This work used the BSD images as they are, without any preprocessing stage.

The segmentation algorithm here presented is compared with the well-known techniques introduced in Section 3.2. Figure 4 offers some examples of the segmentation results obtained with all the algorithms. As it can be seen, there are some "easy" images like the first one and more complicated

---

[1] Note that each measure has different criteria and these criteria are particularly important for understanding the graphical results offered in Section 3.3.

ones like the others. In fact, in this database there are many images that have a rich presence of textures and colours, being these images quite difficult to segment for general-purpose algorithms. Moreover, the second row of the figure also shows four randomly selected examples of the segmentation results produced by humans. By means of this figure, the reader will be able to have a first and coarse idea of the performance of each algorithm.

Table 1 presents the scheme for the experimental part of this work and a brief explanation about the different stages into which the experiments have been divided. Summarising, from each image of the BSD and for each segmentation algorithm, an image segmentation result is produced. On these results, the measures for the segmentation quality are applied, obtaining an explicit value that represents their segmentation goodness in a summary file. This summary file is used to obtain the graphical results and the rankings for the segmentation methods. Each ranking is produced for each quality measure and shows how many times a method was the first, the second and so on. At the end of the process, in order to analyse the statistical significance of these rankings for all the methods used in the comparison, a Friedman test has been performed.

Friedman test [47] is a non-parametric technique to measure the significance of the statistical difference of several methods. These methods provide results on the same problem, using rankings of results obtained by the algorithms to be compared. The Friedman estimator F_F follows a Fisher distribution that allows to analyse the statistical significance of the results. This estimator is expressed as

$$\chi_F^2 = \frac{12N_B}{N_M(N_M+1)} \left(\sum_{j=1}^{N_M} R_j^2 - \frac{N_M(N_M+1)^2}{4} \right)$$

$$F_F = \frac{(N_B-1)\chi_F^2}{N_B(N_M-1) - \chi_F^2}$$

where N_M is the number of segmentation methods, N_B is the number of databases (images) compared and R_j is the average of the ranks for the method j. F_F follows a Fisher distribution with $N_M - 1$ and $(N_M - 1) * (N_B - 1)$ degrees of freedom.

The human segmentation results provided by the BSD will be used as perceptual reference about the goodness of the rest of segmentation results. For each image of the database, h segmentation results made by humans are available, being $5 \leq h \leq 9$ (i.e. at least 5 segmentation results per image). These segmentation results are separately evaluated, obtaining a summary file that will have h columns and one hundred rows. The average value for each row is calculated obtaining a 100-row column that represents the average of the segmentation quality values obtained from the results made by humans. The variance of these segmentation quality values has been also worked out and it was always lower than 0.01, which makes sense to consider only the average value of the human segmentation results. These human average values are

treated as another method in the comparative which will be called **H-avg**.

As it can be found in [46], the BSD subset of one hundred images ranks the images according to some boundary detection benchmarks. In our graphs, images have been re-ordered in a different way depending on the quantitative evaluation obtained by each measure using the **H-avg** values. Thus, a monotonic reference of how the images increase their complexity with regard to the results specified by humans was expected.

Once the human reference has been specified and introduced as another segmentation method in our summary file, Table 2 shows the ranking of methods for the evaluation measures E, E_{CW}, Z_{eb} and F_{RC}. These values represent how many times each method ranks in each position, i.e., first (P1), second (P2),..., seventh (P7). Although some conclusions could be obtained about which are the best results, it is difficult to have a global view of them and, in any case, it would be desirable an analytical measurement of the results. As the scheme of the Table 1 describes, we have selected the Friedman test to this end. This test has been eventually applied to the ranking tables, resulting in a positive evaluation for all of them. For each measure E, E_{CW}, Z_{eb} and F_{RC}, a critical value of the Fisher distribution with a significance level $\alpha = 0.05$, 95% has been set up for the six segmentation algorithms plus the human reference ($N_M = 7$) in the comparative of a hundred images ($N_B = 100$).

As it can be seen, **PSEG** method is 82 times the best according to the E measure, 70 times according to the Z_{eb} measure and 56 times if we use the F_{RC} measure. However, our proposal obtains almost the worst result according to the E_{CW} measure.

Regarding to E and F_{RC} measures, **PSEG** and **H-avg** share the two first positions. Results on these measures have especial meaning since, as it has been said, these measures obtained the best accuracies when machine segmentations were contrasted against the human segmentations in [28]. According to E_{CW}, we obtain very weak results, however, humans also have poor results with this measure. Finally, we obtain the best results with the Z_{eb} measure, nevertheless, human results do not follow this tendency. The three highest values for each method have been written in bold letters in Table 2 in order to show better which is the tendency of each method. Moreover, to support these conclusions, a linear correlation coefficient between each method and **H-avg** has been also worked out for all the quality measures. Thus, **GSEG**, **PSEG** and **JSEG** (in this order) presented the most correlated results with humans in a global sense.

With regard to the global performance demonstrated by each image segmentation algorithm, the method here presented obtains the best results in three out of four quantitative measures for the segmentation goodness. About the perceptual representation obtained by each image segmentation algorithm, our proposal follows the same tendency as humans in three out of four rankings exposed, being its results very correlated with **H-avg**.
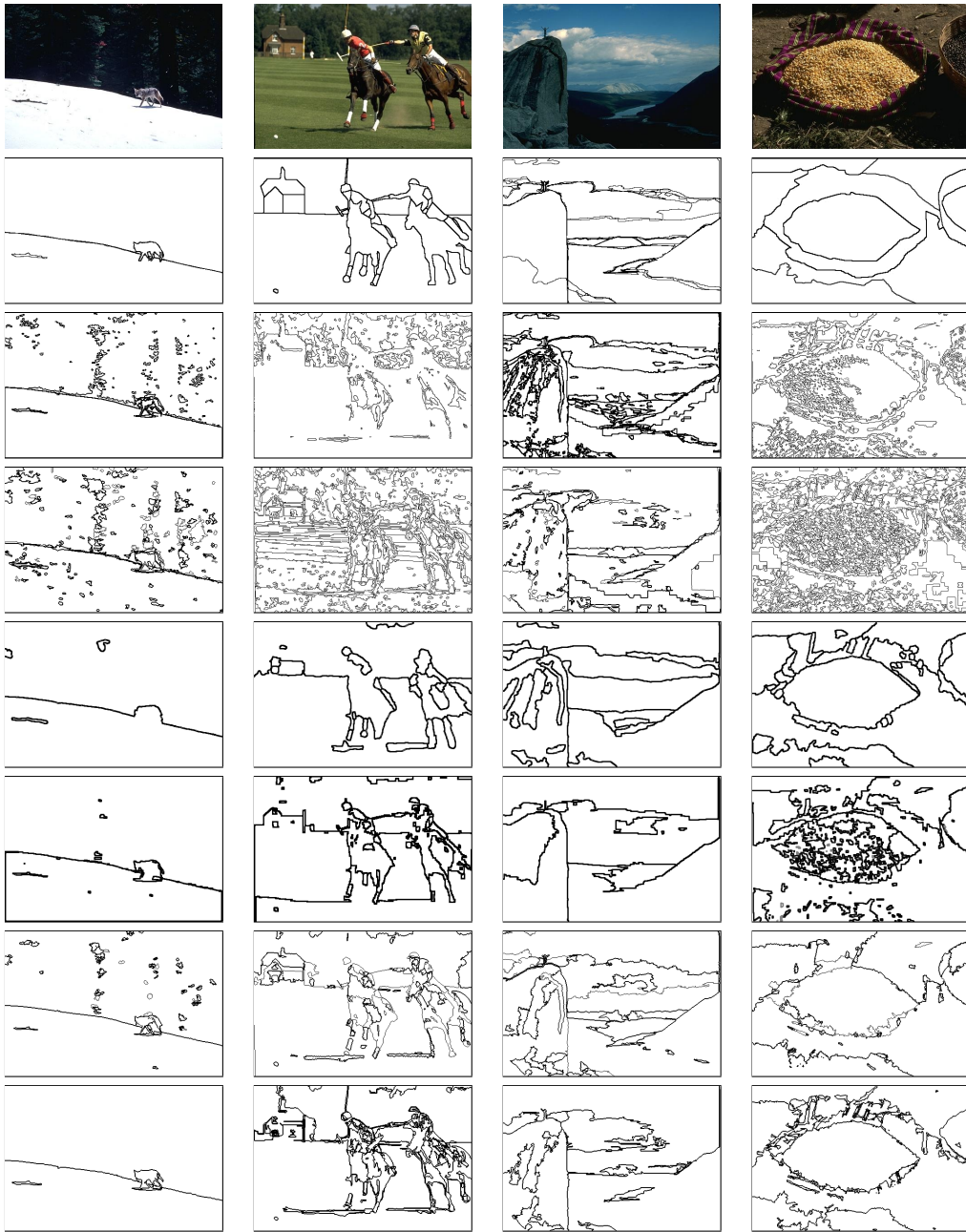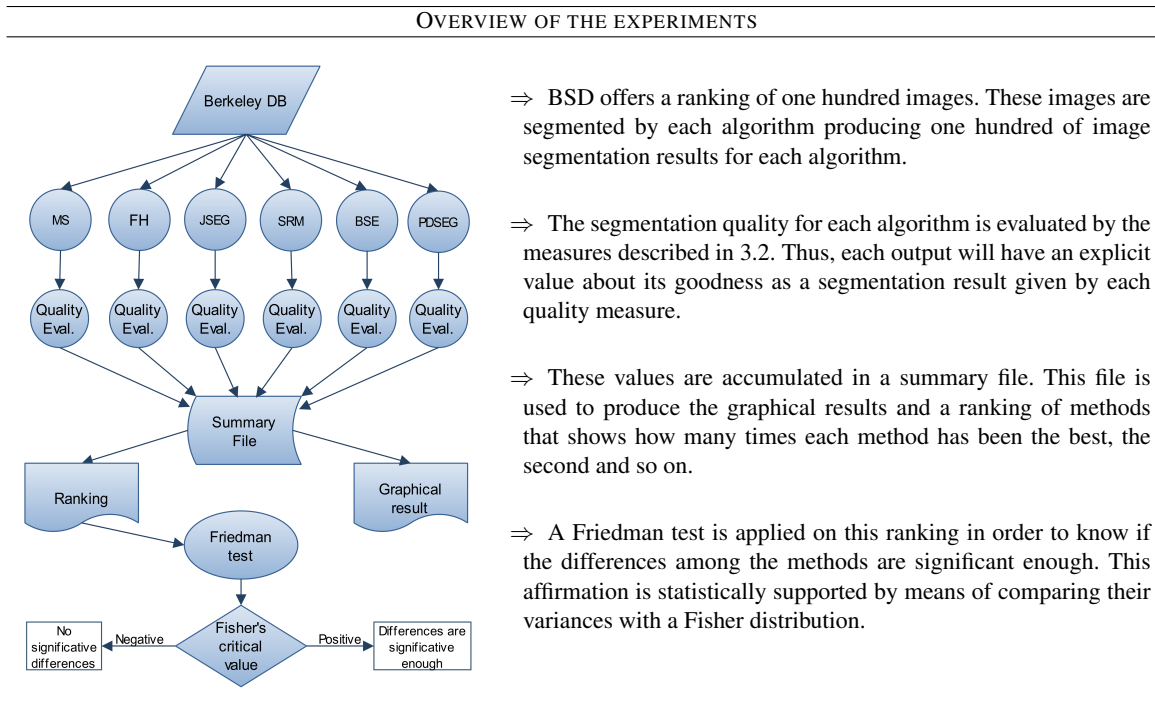
Fig. 4 Segmentation results for test images from the BSD. First row shows the source images and second row shows segmentation results made by humans (randomly chosen). The rest of the rows show the segmentation results for the algorithms **MS**, **FH**, **JSEG**, **SRM**, **GSEG** and **PSEG** (our proposal) respectively.

It is also interesting to notice how some of the segmentation algorithms studied outperform the manually segmented results. This result makes sense because segmentations made by humans are mostly influenced by their prior knowledge about the contents of the image instead of merging those regions that have similar low-level features like texture, colour, etc. Moreover, in our particular case, the proposed algorithm maximises a measure (Eq. (4)) which pursues similar goals than the evaluation measures, that is, to maximize the intra-region homogeneity/uniformity while maximizing the inter-region disparity/contrast as well.

Figure 5 supports this last point showing four different human segmentation images provided by the BSD (second to fifth images). Human segmentation images from BSD have been sorted from left to right by the number of regions in the segmentation result. As it can be seen, these human segmentations are certainly influenced by the semantic contents of the image. The three results on the left (especially the first two) divide the image into amphora and background. These three results will be very penalised by the measures for evaluating the segmentation since humans have not taken into account different tonalities, textures and multiple edges

OVERVIEW OF THE EXPERIMENTS

⇒ BSD offers a ranking of one hundred images. These images are segmented by each algorithm producing one hundred of image segmentation results for each algorithm.

⇒ The segmentation quality for each algorithm is evaluated by the measures described in 3.2. Thus, each output will have an explicit value about its goodness as a segmentation result given by each quality measure.

⇒ These values are accumulated in a summary file. This file is used to produce the graphical results and a ranking of methods that shows how many times each method has been the best, the second and so on.

⇒ A Friedman test is applied on this ranking in order to know if the differences among the methods are significant enough. This affirmation is statistically supported by means of comparing their variances with a Fisher distribution.

Table 1 Experiments overview. Left side shows a flowchart of how the experiments have been developed. Right side offers a brief explanation for each level of the flowchart.

that are present on the surface of the amphora. On the contrary, a good evaluation value will be obtained by the fourth ground-truth image. Obviating the complexity of obtaining such a great segmentation, in this segmentation the regions have been well-separated from a low-level point of view. However, no ground-truth image shows, for instance, the scratched part on the bottom of the amphora where a quite obvious edge can be found and which is quite different in colour to its neighbouring regions.

The main drawback of the rankings is that they do not measure differences among the methods. To solve this point, Figures from 6 to 9 show the quantitative evaluation graphs of each segmentation quality measure (y-axis) with regard to the images of the BSD (x-axis). In these plots, the worst segmentation method according to Table 2 has not been included, that is, the **FH** algorithm in E, E_{CW} and F_{RC} measures and the **JSEG** algorithm in Z_{eb} measure. In addition, graphical results for each evaluation measure are shown in two plots. Again according to the results shown in Table 2, the plot on the top shows the two best algorithms for each evaluation measure whereas the one on the bottom shows the rest of algorithms[2]. Presenting the results of all the algorithms in the same plot results in graphs that are almost impossible to understand. Thus, by means of separating the plots in this way, an improved view of the results is expected. Also, remember that, in these plots, the BSD images have been re-ordered according to the quantitative evaluation values of the **H-avg** for each evaluation measure. Therefore, image numbers (x-

axis) in figures do not correspond to the same image for each quantitative evaluation measure.

Graphical results confirm the ranking results. The proposed **PSEG** algorithm obtains the best results by far for E and F_{RC} measures. We are still better using the Z_{eb} measure, although the differences with the **SRM** algorithm are not so evident in this case. Regarding to the E_{CW} measure, as well as it happened with the **H-avg** results, our proposal obtains the worst results in most of the images.

Finally, it is important to notice that no monotonic graph was reached using the order provided by the BSD ranking of images. By re-ordering the images according to the evaluation of the segmentation results produced by humans, a monotonic reference of how the images increase their segmentation difficulty was again expected. E_{cw} and Z_{eb} measures seem to have a global monotonic behaviour, however, even using this new order, we obtained similar saw tooth graphs in all the cases. In addition, we have confirmed that, independently from the reference used for sorting the images of the database, the rest of the methods will draw a graph with multiple ups and downs. This point lead us to think that there exists a weak relationship among the image segmentation methods and the unsupervised evaluation measures for judging the quality of the segmentation.

4 Conclusions and future work

An unsupervised colour segmentation algorithm has been presented in this paper. The proposed method performs a hierarchical structure of partitions by means of a region merging

---

[2] The same scale for y-axes has been applied in both plots.
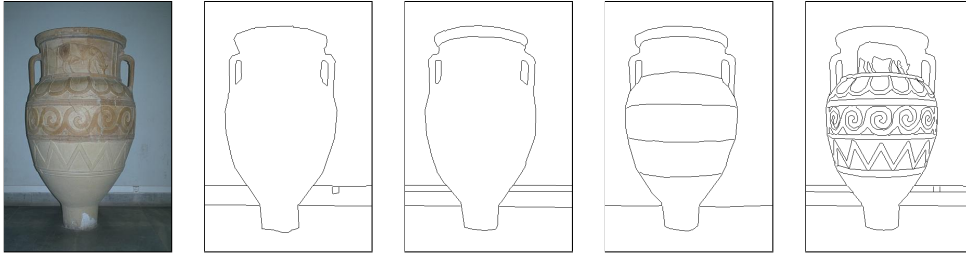
Fig. 5 Segmentation results made by humans from the BSD. First image on the left shows the source image. Second to fifth images are manually labelled results.

E MEASURE

	P1	P2	P3	P4	P5	P6	P7
MS	0	2	4	**13**	**25**	**56**	0
FH	0	0	0	0	0	0	**100**
JSEG	0	15	13	**28**	**29**	**15**	0
SRM	0	5	20	**28**	**21**	**26**	0
GSEG	2	**32**	**33**	**18**	15	0	0
PSEG	**82**	**6**	**4**	2	3	3	0
H-avg	**16**	**40**	**26**	11	7	0	0

E_{CW} MEASURE

	P1	P2	P3	P4	P5	P6	P7
MS	**24**	15	14	12	**17**	**17**	1
FH	0	1	0	0	**4**	**16**	**79**
JSEG	**59**	**33**	**6**	2	0	0	0
SRM	14	**27**	**34**	**16**	6	3	0
GSEG	3	**18**	**36**	**29**	12	2	0
PSEG	0	2	1	8	**27**	**44**	**18**
H-avg	0	4	9	**33**	**34**	**18**	2

Z_{eb} MEASURE

	P1	P2	P3	P4	P5	P6	P7
MS	4	13	13	**20**	**20**	11	**19**
FH	**16**	13	**25**	10	12	4	**20**
JSEG	0	1	3	9	**21**	**31**	**35**
SRM	8	**41**	**32**	**11**	6	2	0
GSEG	0	2	8	**37**	**17**	**29**	7
PSEG	**70**	**25**	**5**	0	0	0	0
H-avg	2	5	14	13	**24**	**23**	**19**

F_{RC} MEASURE

	P1	P2	P3	P4	P5	P6	P7
MS	13	13	**15**	**26**	**16**	15	2
FH	0	5	9	5	**22**	**30**	**29**
JSEG	1	7	9	**18**	17	**21**	**27**
SRM	2	4	7	**20**	**27**	18	**22**
GSEG	9	**14**	**41**	**20**	9	4	3
PSEG	**56**	**12**	3	6	5	**9**	9
H-avg	**19**	**45**	**16**	5	4	3	8

Table 2 Ranking of algorithms for E, E_{CW}, Z_{eb} and F_{RC} measures. The three highest values for each method are in bold letters.

Fig. 6 Segmentation quality measure E (y-axis). Images of the BSD (x-axis) are *increasingly* sorted according to the quality value obtained by **H-avg**: Plot on the top for **GSEG** and **PSEG** algorithms. Plot on the bottom for **MS**, **JSEG** and **SRM** algorithms. The *lower* the value of the measure, the better the segmentation results.

procedure. The whole approach has demonstrated a good behaviour mainly due to *i)* the proposed criterion for evaluating how suitable each partition is according to the low perceptual contents of the scene and *ii)* the measure of similarity among regions used as a merging criterion during the agglomerative process. In order to check the relevance of the proposed solution, it was widely tested on a database of real images and compared against some of the most referenced methods in the colour image segmentation literature. In addition, the results

produced in each case were evaluated by means of unsupervised methods for measuring the quality of the segmentation results.

In this work, manual segmentations specified by humans are assumed as the best perceptual reference. Although these manual segmentations show a content-based image segmentation with high-level semantics, there also exists an important low-level perceptual basis in these segmentation results. Under this premise, the experimental part of this work shows how the proposed algorithm finds low-level perceptual re-
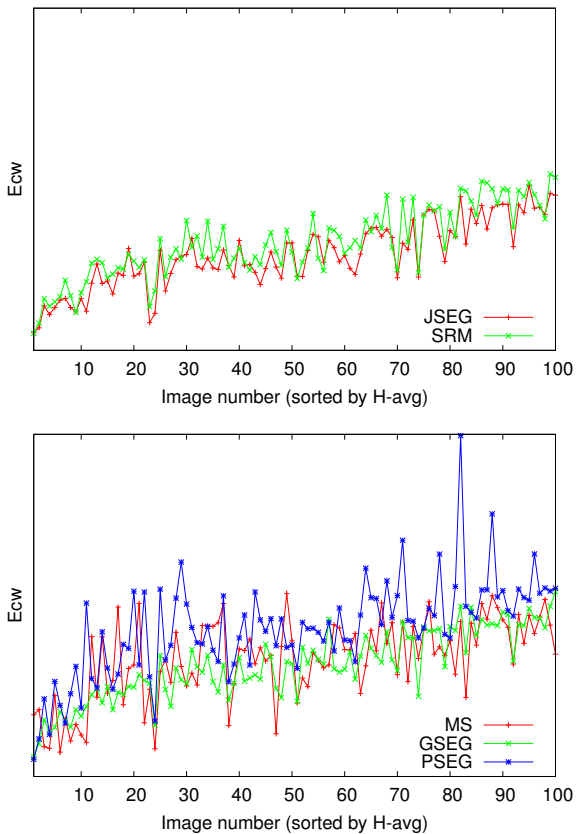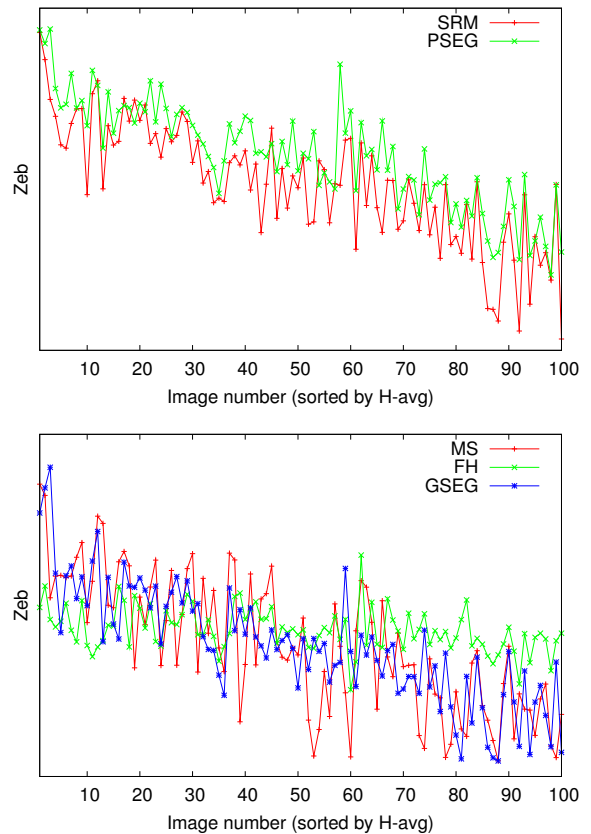
Fig. 7 Segmentation quality measure E_{CW} (y-axis). Images of the BSD (x-axis) are *increasingly* sorted according to the quality value obtained by **H-avg**: Plot on the top for **JSEG** and **SRM** algorithms. Plot on the bottom for **MS**, **GSEG** and **PSEG** algorithms. The *lower* the value of the measure, the better the segmentation results.
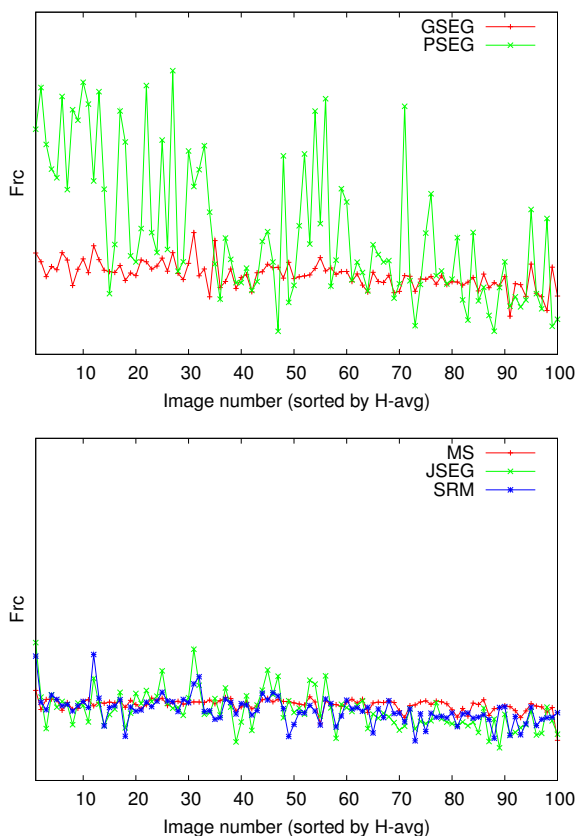
Fig. 8 Segmentation quality measure Z_{eb} (y-axis). Images of the BSD (x-axis) are *decreasingly* sorted according to the quality value obtained by **H-avg**: Plot on the top for **SRM** and **PSEG** algorithms. Plot on the bottom for **MS**, **GSEG** and **PSEG** algorithms. The *higher* the value of the measure, the better the segmentation results.

gions in a very similar way to humans. Thus, we obtained similar patterns of behaviour to humans in 3 out of 4 measures. The primary regions obtained can be used as input for higher semantic-based segmentation processes.

Instead of using a single measure for quantifying the segmentation quality, some authors support the idea of taking a collection of similar measures for defining an overall performance measure [48]. In a global sense, our proposal has reached the best performance in terms of the quantitative measures for the segmentation goodness, obtaining the best results in three out of four of them.

It is also interesting to discuss separately from our proposal about the performance shown by the rest of the image segmentation methods since they have widely been used as a reference in other comparisons. Thus, **GSEG** has demonstrated to be the most correlated algorithm with humans. **MS** and **FH** algorithms are probably the most used methods for comparative purposes. However, both algorithms, and especially **FH**, have demonstrated a surprisingly poor performance in comparison with the rest of the methods. **JSEG** and **SRM** algorithms have reached a reasonable performance, being **JSEG** the best one when the E_{CW} measure was used for quantifying the quality of the solutions obtained by the algorithms.

Many well-known publications support the use of unsupervised measures as a way of judging the segmentation quality [27, 42, 28] and they have been generally accepted in the scientific community. However, we have observed some inconsistencies on these measures when results from the segmentation algorithms are compared to the results provided by humans. In [28], authors warn about how some evaluation methodologies follow approximately the same criteria as some segmentation algorithms. This point may lead us to wrongly believe that these algorithms are much better than they actually are, being even more accurate than the human segmentation reference. It is especially interesting to observe this point in the results presented in this work where human quantitative values about the segmentation quality are often beaten. This fact should be, at least semantically, considered as a wrong result for these quality segmentation measures.

The key idea of this paradoxical behaviour could probably arise from considering image segmentation procedures as high-level tasks. For this kind of tasks, human references would be undoubtedly better. However, if we consider image segmentation as a low-level process, it is certainly possible to achieve better results than the references provided by humans.

Fig. 9 Segmentation quality measure F_{RC} (y-axis). Images of the BSD (x-axis) are *decreasingly* sorted according to the quality value obtained by **H-avg**: Plot on the top for **GSEG** and **PSEG** algorithms. Plot on the bottom for **MS**, **JSEG** and **SRM** algorithms. The *higher* the value of the measure, the better the segmentation results.

Traditionally, unsupervised image segmentation has been considered as a low-level process, however, nowadays there is no doubt that the image segmentation results are mostly evaluated according to their semantic contents. Although in most applications the results need to be as close as possible to the human behaviour, this semantic image segmentation gives rise to an ill-posed problem and, in this way, no unsupervised algorithm would be able to reach that level of abstraction.

From our point of view, image segmentation in an unsupervised way may only have sense as a low-level procedure which, in this sense, can be evaluated by quantitative measures for the segmentation quality. Thus, lower perceptual stages could be somehow approximated by means of imitating the organisation processes followed by humans. As far as we concern, our proposal achieves this approximation, improving the state-of-the-art in this direction.

The main drawback of the proposed image segmentation process is probably its computational cost. Although, no objective comparison has been possible[3], the other algorithms that participated in the comparative are definitely faster than

our proposal. It is worth saying that our particular implementation is not optimised. As a future work, the computational cost is expected to be reduced when the algorithm is adapted to the particular purposes of an application domain. Thus, if some a priori information or knowledge about the system is incorporated in a *semi-supervised segmentation* process, the algorithm will necessarily reduce its requirements, and thus optimise the segmentation results according to the needs of each application.

5 Biographies

References

1. Pal, N. R., Pal, S. K., 1993. A review on image segmentation techniques. Pattern Recognition 26 (9), 1277–1294.
2. Zucker, S., 1976. Region growing: Childhood and adolescence. CGIP 5, 382–399.
3. Fu, K., Mui, J., 1981. A survey on image segmentation. Pattern Recognition 13, 3–16.
4. Lucchese, L., Mitra, S., 1999. Advances in color image segmentation. GLOBECOM 4, 2038–2044.
5. Haralick, R. H., Shapiro, L. G., 1985. Image segmentation techniques. CVGIP 29, 100–132.
6. Cheng, H.-D., Jiang, X. H., Sun, Y., Wang, J., 2001. Color image segmentation: Advances and prospects. Pattern Recognition 34 (12), 2259–2281.
7. Sahoo, P. K., Soltani, S., Wong, A. K., Chen, Y. C., 1988. A survey of thresholding techniques. Comput. Vision Graph. Image Process. 41 (2), 233–260.
8. Tabb, M., Ahuja, N., 1997. Multiscale image segmentation by integrated edge and region detection. IEEE Trans on Image Processing 6 (5), 642–655.
9. Todorovic, S., Ahuja, N., 2008. Unsupervised category modeling, recognition and segmentation in images. IEEE Trans on PAMI 30 (12), 2158–2174.
10. Beveridge, J., Griffith, J. S., Kohler, R. R., Hanson, A., Riseman, E., 1989. Segmenting images using localized histograms and region merging. Int J of Computer Vision 2 (3), 311–352.
11. Rubner, Y., Puzicha, J., Tomasi, C., Buhmann, J. M., 2001. Empirical evaluation of dissimilarity measures for color and texture. CVIU 84(1), 25–43.

---

[3] Image results for GSEG algorithm were provided by the author for all the BSD and they offer an average time of 24sec. per image in their paper. Likewise, our implementation and the ones for FH, MS

---

and JSEG algorithms have been programmed in C language, however, the SRM algorithm were provided in MATLAB by the authors.

12. Todorovic, S., Ahuja, N., 2006. Extracting subimages of an unknown category from a set of images. CVPR. pp. 927–934.

13. Tremeau, A., Borel, N., 1997. A region growing and merging algorithm to color segmentation. Pattern Recognition 30 (7), 1191–1203.

14. Pauwels, E. J., Frederix, G., 1999. Finding salient regions in images: Non-parametric clustering for image segmentation and grouping. CVIU 75(1/2), 73–85.

15. Randall, J., Guan, L., Li, W., Zhang, X., 2008. The HCM for perceptual image segmentation. Neurocomputing 71 (10–12), 1966–1979.

16. Haxhimusa, Y., Kropatsch, W. G., 2004. Segmentation graph hierarchies. In Proceedings of the SSPR-SPR, 343–351.

17. Mirmehdi, M., Petrou, M., 2000. Segmentation of color textures. IEEE Trans on PAMI 22 (2), 142–159.

18. Chen, J., Pappas, T. N., Mojsilović, A., Rogowitz, B. E., 2005. Adaptive perceptual color-texture image segmentation. IEEE Trans on Image Processing 14 (10), 1524–1536.

19. Shi, J., Malik, J., 2000. Normalized cuts and image segmentation. IEEE Trans on PAMI 22 (8), 888–905.

20. Gdalyahu, Y., Weinshall, D., Werman, M., 2001. Self-organization in vision: Stochastic clustering for image segmentation, perceptual grouping, and image database organization. IEEE Trans on PAMI 23, 10531074.

21. Paschos, G., 2001. Perceptually uniform color spaces for color texture analysis: an empirical evaluation. IEEE Trans on Image Processing 10 (6), 932–937.

22. Alata, O., Quintard, L., 2009. Is there a best color space for color image characterization or representation based on Multivariate Gaussian Mixture Model?. Computer Vision and Image Understanding 113 (8), 867–877.

23. Zhang, H., Goldman, S. A., 2006. Perceptual information of images and the bias in homogeneity-based segmentation. In Proceedings of the CVPR, pp. 181–188.

24. Palmer, S., Rock, I., 1994. Rethinking perceptual organization: the role of uniform connectedness. Psychonomic Bulletin and Review (1), 29–55.

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

26. Cardoso, J. S., Corte-Real, L., 2005. Toward a generic evaluation of image segmentation. IEEE Trans on Image Processing 14 (11), 1773– 1782.

27. Chabrier, S., Emile, B., Laurent, H., Rosenberger, C., Marché, P., 2004. Unsupervised evaluation of image segmentation application to multi-spectral images. ICPR 1, 576–579.

28. Zhang, H., Fritts, J. E., Goldman, S. A., 2008. Image segmentation evaluation: A survey of unsupervised methods. Computer Vision and Image Understanding 110 (2), 260–280.

29. Deng, Y., Kenney, C., Moore, M.S., Manjunath, B.S., 1999. Peer group filtering and perceptual color image quantization. In Proceedings of the IEEE ISCS vol.4, 21–24.

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

31. Weiss, Y., Adelson, E. H., 1996. A unified mixture framework for motion segmentation: incorporating spatial coherence and estimating the number of models. CVPR. pp. 321-326.

32. Mansouri, A-R., Mitiche, A., Vázquez, C., 2006. Multiregion competition: A level set extension of region competition to multiple region image partitioning. Computer Vision and Image Understanding 101 (3), 137–150.

33. Mohand Saïd Allili, Nizar Bouguila, Djemel Ziou, 2008. Finite general Gaussian mixture modeling and application to image and video foreground segmentation. J. Electronic Imaging., vol.17(1), 013005.

34. Mohand Saïd Allili, Djemel Ziou, Nizar Bouguila, Sabri Boutemedjet, 2010. Image and Video Segmentation by Combining Unsupervised Generalized Gaussian Mixture Modeling and Feature Selection. IEEE Trans. Circuits Syst. Video Techn., vol.20(10), 1373–1377.

35. Comaniciu, D., Meer, P., 1997. Robust analysis of feature spaces: Color image segmentation. IEEE Conf. Computer Vision and Pattern Recognition, 750–755.

36. Felzenszwalb, P. F., Huttenlocher, D. P., 2004. Efficient graph-based image segmentation. International Journal of Computer Vision 59 (2), 167–181.

37. Nock, R., Nielsen, F., 2004. Statistical region merging. IEEE Trans on PAMI 26 (11), 1452–1458.

38. Deng, Y., Manjunath, B., August 2001. Unsupervised segmentation of color-texture regions in images and video. IEEE Trans on PAMI 23 (8), 800–810.

39. Ugarriza, L. G., Saber, E., Vantaram, S. R., Amuso, V., Shaw, M., Bhaskar, R., 2009. Automatic image segmentation by dynamic region growth and multiresolution merging. IEEE Trans on Image Processing 18 (10), 2275–2288.

40. Martinez-Uso, A., Pla, F., Garcia-Sevilla, P., 2006. Unsupervised image segmentation using a hierarchical clustering selection process. LNCS (4109), 799–807.

41. Unnikrishnan, R., Pantofaru, C., Hebert, M., 2007. Toward objective evaluation of image segmentation algorithms. IEEE Trans on PAMI 29 (6), 929–944.

42. Zhang, H., Fritts, J. E., Goldman, S. A., 2004. An entropy-based objective evaluation method for image segmentation. pp. In Proceedings of the SPIE, 38–49.

43. Chen, H.-C., Wang, S.-J., 2004. The use of visible color difference in the quantitative evaluation of color image segmentation. In: IEEE ICASSP. Vol. 3. pp. 593–596.

44. Zeboudj, R., 1998. Filtrage, seuillage automatique, contraste et contours: du pré-traitement à l'analyse d'image. Ph.D. thesis, University of Saint Etienne, France.

45. Rosenberger, C., Chehdi, K., 2000. Genetic fusion: application to multi-components image segmentation. In: IEEE ICASSP. pp. 2223–2226.

46. Martin, D., Fowlkes, C., Tal, D., Malik, J., 2001. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In: In Proceedings of 8th ICCV. Vol. 2. pp. 416–423.

47. Friedman, M., 1937. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. American Statistical Association 32 (200), 675–701.

48. Jiang, X., Marti, C., Irniger, C., Bunke, H., 2006. Distance measures for image segmentation evaluation. EURASIP J. Appl. Signal Process., 1–10.