A New Perception-Based Segmentation Approach Using Combinatorial Pyramids

Esther Antúnez, Rebeca Marfil, and Antonio Bandera

Grupo ISIS, Dpto. Tecnología Electrónica, ESTI Telecomunicación,
Universidad de Málaga
Campus de Teatinos, 29071-Málaga, Spain
{eantunez,rebeca,ajbandera}@uma.es

Abstract. This paper presents a bottom-up approach for perceptual segmentation of natural images. The segmentation algorithm consists of two consecutive stages: firstly, the input image is partitioned into a set of blobs of uniform colour (pre-segmentation stage) and then, using a more complex distance which integrates edge and region descriptors, these blobs are hierarchically merged (perceptual grouping). Both stages are addressed using the Combinatorial Pyramid, a hierarchical structure which can correctly encode relationships among image regions at upper levels. Thus, unlike other methods, the topology of the image is preserved. The performance of the proposed approach has been initially evaluated with respect to groundtruth segmentation data using the Berkeley Segmentation Dataset and Benchmark. Although additional descriptors must be added to deal with textured surfaces, experimental results reveal that the proposed perceptual grouping provides satisfactory scores.

Keywords: perceptual grouping, irregular pyramids, combinatorial pyramids.

1 Introduction

Image segmentation is the process of decomposing an image into a set of regions which have some similar visual characteristics. These visual characteristics can be based on pixel properties as color, brightness or intensity or on other more general properties as texture or motion. Segmentation in regions may be achieved using pyramidal methods that provide hierarchical partitions of the original image. These pyramidal structures help in reducing the computational load associated to the segmentation process and allows to have a same object in different levels of representation. Basically, a pyramid represents an image at different resolution levels. Each pyramid level is recursively obtained by processing its underlying level. In this hierarchy, the bottom level contains the image to be processed. The main advantage of the pyramidal structure is that the parent-child relationships defined between nodes in adjacent levels can be used to reduce the time required to analyze an image. Besides, among the inherent properties of pyramids are [3]: reduction of noise and computational cost, resolution independent processing, processing with local and global features within the

G. Maino and G.L. Foresti (Eds.): ICIAP 2011, Part I, LNCS 6978, pp. 327–336, 2011.

same frame. Moreover, irregular pyramids adapt their structure to the data. A detailed explanation of pyramidal structures can be found in [12]. Combinatorial Pyramids are irregular pyramids in which each level of the pyramid is defined by a combinatorial map. A combinatorial map is a mathematical model describing the subdivision of a space. It encodes all the vertices which compound this subdivision and all the incidence and adjacency relationships among them. In this way, the topology of the space is fully described.

On the other hand, natural images are generally composed of physically disjoint objects whose associated groups of image pixels may not be visually uniform. Hence, it is very difficult to formulate a priori what should be recovered as a region from an image or to separate complex objects from a natural scene [10]. To achieve this goal several authors have proposed generic segmentation methods called 'perceptual segmentations', which try to divide the input image in a manner similar to human beings. Therefore, perceptual grouping can be defined as the process which allows to organize low-level image features into higher level relational structures. Handling such high-level features instead of image pixels offers several advantages such as the reduction of computational complexity of further processes. It also provides an intermediate level of description (shape, spatial relationships) for data, which is more suitable for object recognition tasks [16].

As the process to group pixels into higher level structures can be computationally complex, perceptual segmentation approaches typically combine a pre-segmentation stage with a subsequent perceptual grouping stage [1]. The pre-segmentation stage conducts the low-level definition of segmentation as a process of grouping pixels into homogeneous clusters, meanwhile the perceptual grouping stage performs a domain-independent grouping which is mainly based on properties such as the proximity, similarity, closure or continuity. It must be noted that the aim of these approaches is providing a mid-level segmentation which is more coherent with the human-based image decomposition. That is, it could be usual that the final regions obtained by these bottom-up approaches do not always correspond to the natural image objects [8,13].

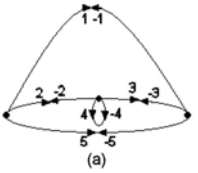
This paper presents a hierarchical perceptual segmentation approach which accomplishes these two aforementioned stages. The pre-segmentation stage uses a colour-based distance to divide the image into a set of regions whole spatial distribution is physically representative of the image content. The aim of this stage is to represent the image by means of a set of blobs (superpixels) whose number will be commonly very much less than the original number of image pixels. Besides, these blobs will preserve the image geometric structure as each significant feature contain at least one blob. Next, the perceptual grouping stage groups this set of homogeneous blobs into a smaller set of regions taking into account not only the internal visual coherence of the obtained regions but also the external relationships among them. Both stages are addressed using the Combinatorial Pyramid. It can be noted that this framework is closely related to the previous works of Arbeláez and Cohen [1,2], Huart and Bertolino [8] and Marfil and Bandera [11]. In all these proposals, a pre-segmentation stage precedes the

perceptual grouping stage: Arbeláez and Cohen propose to employ the extrema mosaic technique [2], Huart and Bertolino use the Localized Pyramid [8] and Marfil and Bandera employ the Bounded Irregular Pyramid (BIP) [11]. The result of this first grouping is considered in all these works as a graph, and the perceptual grouping is then achieved by means of a hierarchical process whose aim is to reduce the number of vertices of this graph. Vertices of the uppermost level will define a partition of the input image into a set of perceptually relevant regions. Different metrics and strategies have been proposed to address this second stage, but all of the previously proposed methods rely on the use of a simple graph (i.e., a region adjacency graph (RAG)) to represent each level of the hierarchy. RAGs have two main drawbacks for image processing tasks: (i) they do not permit to know if two adjacent regions have one or more common boundaries, and (ii) they do not allow to differentiate an adjacency relationship between two regions from an inclusion relationship. That is, the use of this graph encoding avoids that the topology will be preserved at upper levels of the hierarchies. Taking into account that objects are not only characterized by features or parts, but also by the spatial relationships among these features or parts [15], this limitation constitutes a severe disadvantage. Instead of simple graphs, each level of the hierarchy could be represented using a dual graph. Dual graphs preserve the topology information at upper levels representing each level of the pyramid as a dual pair of graphs and computing contraction and removal operations within them [9]. Thus, they solve the drawbacks of the RAG approach. The problem of this structure is the high increase of memory requirements and execution times since two data structures need now to be stored and processed. Combinatorial maps can be seen as an efficient representation of dual graphs in which the orientation of edges around the graph vertices is explicitly encoded using only one structure. Thus, the use of this structure reduces the memory requirements and execution times.

The rest of the paper is organized as follows: Section 2 describes the proposed approach. It briefly explains the main aspects of the pre-segmentation and perceptual grouping processes which are achieved using the Combinatorial Pyramid. Experimental results revealing the efficiency of the proposed method are presented in Section 3. Finally, the paper concludes along with discussions and future work in Section 4.

2 Segmentation Algorithm

As we aforementioned, the perceptual segmentation algorithm is divided in two stages: pre-segmentation and perceptual grouping stages. Moreover, in both stages the combinatorial map is employed as the data structure to represent each level of the pyramid. Combinatorial maps define a general framework, which allows to encode any subdivision of nD topological spaces orientable or non-orientable with or without boundaries. Formally speaking, a n-dimensional combinatorial map is described as a $(n+1)$-tuple $M = (D, \beta_1, \beta_2, ..., \beta_n)$ such that D is the set of abstract elements called *darts*, β_1 is a permutation on D



dart	1	-1	2	-2	3	-3	4	-4	5	-5
α	-1	1	-2	2	-3	3	-4	4	-5	5
σ	5	-3	1	4	-2	-5	-4	3	2	-1

(b)

Fig. 1. a) Example of combinatorial map; and b) values of α and σ for the combinatorial map in a)

and the other β_i are involutions on D. An involution is a permutation whose cycle has the length of two or less.

In the case of 2D, combinatorial maps may be defined with the triplet $G = (D, \alpha, \sigma)$, where D is the set of darts, σ is a permutation in D encoding the set of darts encountered when turning (counter) clockwise around a vertex, and α is an involution in D connecting two darts belonging to the same edge:

$$\forall d \in D, \alpha^2(d) = d \tag{1}$$

Fig. 1.a) shows an example of combinatorial map. In Fig. 1.b) the values of α and σ for such a combinatorial map can be found. In our approach, counter-clockwise orientation (ccw) for σ is chosen.

The symbols $\sigma^*(d)$ and $\alpha^*(d)$ stand, respectively, the σ and α orbits of the dart d. The orbit of a permutation is obtained applying successively such a permutation over the element that is defined. In this case, the orbit σ^* encodes the set of darts encountered when turning counter-clockwise around the vertex encoded by the dart d. The orbit α^* encode the darts that belong to the same edge. Therefore, the orbits of σ encode the vertices of the graph and the orbits of α define the edges of the graph. In the example of Fig. 1, $\alpha^*(1) = \{1, -1\}$ and $\sigma^*(1) = \{1, 5, 2\}$.

Given a combinatorial map, its dual is defined by $\bar{G} = (D, \varphi, \alpha)$ with $\varphi = \sigma \circ \alpha$. The orbits of φ encode the faces of the combinatorial map. Thus, the orbit φ^* can be seen as the set of darts obtained when turning-clockwise a face of the map. In the example of Fig. 1, $\phi^*(1) = \{1, -3, -2\}$.

Thus, 2D combinatorial maps encode a subdivision of a 2D space into vertices ($\mathrm{V} = \sigma^*(D)$), edges ($\mathrm{E} = \alpha^*(D)$) and faces($\mathrm{F} = \varphi^*(D)$).

When a combinatorial map is built from an image, the vertices of such a map G could be used to represent the pixels (regions) of the image. Then, in its dual $\bar{G}$, instead of vertices, faces are used to represent pixels (regions). Both maps store the same information and there is not so much difference in working with G or $\bar{G}$. However, as the base entity of the combinatorial map is the dart, it is not possible that this map contains only one vertex and no edges. Therefore, if we choose to work with G, and taking into account that the map could be composed by an unique region, it is necessary to add special darts to represent the infinite region which surrounds the image (the background). Adding these darts, it is avoided that the map will contain only one vertex. On the other hand,

when $\bar{G}$ is chosen, the background also exists but there is no need to add special darts to represent it. In this case, a map with only one region (face) would be made out of two darts related by α and σ.

In our case, the base level of the pyramid will be a combinatorial map where each face represent a pixel of the image as an homogeneous region. The combinatorial pyramid is build reducing this initial combinatorial map successively by a sequence of contraction or removal operations [5,9].

In the following subsections, the application of the Combinatorial Pyramid to the pre-segmentation and perceptual grouping stages is explained in detail.

2.1 Pre-segmentation Stage

Let $G_0 = (D_0, \sigma_0, \alpha_0)$ be a given attributed combinatorial map with the vertex set $V_0 = \sigma^*(D)$, the edge set $E_0 = \alpha^*(D)$ and face set $F_0 = \varphi^*(D)$ on the base level (level 0) of the pyramid. In the same way, the combinatorial map on level k of the pyramid is denoted by $G_k = (D_k, \sigma_k, \alpha_k)$. As we aforementioned, each face of the base level represent a pixel of the image. Thus, faces are attributed with the colour of the corresponding pixel. The colour space used in our approach is the HSV space. The edges of the map are also attributed with the difference of colour of the regions separated by each edge. The hierarchy of graphs is built using the algorithm proposed by Haxhimusa et al [7,6], which is based on a spanning tree of the initial graph obtained using the algorithm of Borůvka [4]. Building the spanning tree allows to find the region borders quickly and effortlessly based on local differences in a color space. For each face $f \in F_k$ Borůvka's algorithm marks the edge $e \in E_k$ with the smallest attribute value to be removed. Now, unlike [7,6], two regions (faces) are merged if the difference of colour between them is smaller than a given threshold U_p. That is, the attribute of each edge marked to be removed for the Borůvka's algorithm is compared with the threshold U_p and if its value is smaller, that edge if added to a removal kernel ($RK_{k,k+1}$). In a second step, hanging edges are removed. Finally, a contraction kernel ($CK_{k,k+1}$) is applied to remove parallel edges, obtaining the new level of the pyramid. After a contraction step, the attributes of the surviving edges have to be updated with the colour distance of the faces that the new edge separates. This process is iteratively repeated until no more removal/contraction operation is possible. This stage results in an over-segmentation of the image into a set of regions with homogeneous colour. These homogeneous regions will be the input of the perceptual grouping stage.

2.2 Perceptual Grouping Stage

After the pre-segmentation stage, the perceptual grouping stage aims for simplifying the content of the obtained colour-based image partition. To achieve an efficient grouping process, the Combinatorial Pyramid ensures that two constraints are respected: (i) although all groupings are tested, only the best groupings are locally retained; and (ii) all the groupings are spread on the image so that no part of the image is advantaged. To join pre-segmentation and perceptual

grouping stages, the last level of the Combinatorial Pyramid associated to the pre-segmentation stage will constitute the first level of the pyramid associated to the perceptual grouping stage. Next, successive levels will be built using the decimation scheme described in Section 2.1. However, in order to accomplish the perceptual grouping process, a distance which integrates boundary and region descriptors has been defined as a criteria to merge two faces of the combinatorial map.

The distance has two main components: the colour contrast between image blobs and the boundaries of the original image computed using the Canny detector. In order to speed up the process, a global contrast measure is used instead of a local one. It allows to work with the faces of the current working level, increasing the computational speed. This contrast measure is complemented with internal region properties and with attributes of the boundary shared by both regions. The distance between two regions (faces) $\mathbf{f}_i \in F_k$ and $\mathbf{f}_j \in F_k$, $\psi^{\alpha,\beta}(\mathbf{f}_i, \mathbf{f}_j)$, is defined as

$$\psi^{\alpha,\beta}(\mathbf{f}_i, \mathbf{f}_j) = \frac{d(\mathbf{f}_i, \mathbf{f}_j) \cdot b_{\mathbf{f}_i}}{\alpha \cdot (c_{\mathbf{f}_i\mathbf{f}_j}) + (\beta \cdot (b_{\mathbf{f}_i\mathbf{f}_j} - c_{\mathbf{f}_i\mathbf{f}_j}))} \tag{2}$$

where $d(\mathbf{f}_i, \mathbf{f}_j)$ is the colour distance between $\mathbf{f}_i$ and $\mathbf{f}_j$. $b_{\mathbf{f}_i}$ is the perimeter of $\mathbf{f}_i$, $b_{\mathbf{f}_i\mathbf{f}_j}$ is the number of pixels in the common boundary between $\mathbf{f}_i$ and $\mathbf{f}_j$ and $c_{\mathbf{f}_i\mathbf{f}_j}$ is the set of pixels in the common boundary which corresponds to pixels of the boundary detected by the Canny detector. α and β are two constant values used to control the influence of the Canny boundaries in the grouping process. Two regions (faces) will be merged if that distance, $\psi^{\alpha,\beta}(\cdot,\cdot)$, is smaller than a given threshold U_s. It must be noted that the distance $\psi^{\alpha,\beta}(\cdot,\cdot)$ between two regions (faces) is proportional to its colour distance. However, it must be also noted that this distance decreases if the most of the boundary pixels of one of the regions is in contact with the boundary pixels of the other one. Besides, the distance value will decrease if these shared boundary pixels are not detected by the Canny detector.

3 Experimental Results

In order to evaluate the performance of the proposed colour image segmentation approach, the Berkeley Segmentation Dataset and Benchmark (BSDB) has been employed[1] [14]. In this dataset, the ground-truth data is provided by a large database of natural images, manually segmented by human subjects. The methodology for evaluating the performance of segmentation techniques is based in the comparison of machine detected boundaries with respect to human-marked boundaries using the *Precision-Recall framework* [13]. This technique considers two quality measures: precision and recall. The *precision* (P) is defined as the fraction of boundary detections that are true positives rather than false positives. Thus, it quantifies the amount of noise in the output of the boundaries detector approach. On the other hand, the *recall* (R) is defined by the fraction of true positives that are detected rather than missed. Then, it quantifies

[1] http://www.cs.berkeley.edu/projects/vision/grouping/segbench/

the amount of ground truth detected. Measuring these descriptors over a set of images for different thresholds of the approach provides a parametric Precision-Recall curve. The F-measure combines these two quality measures into a single one. It is defined as their harmonic mean:

$$F(P, R) = \frac{2PR}{P + R} \tag{3}$$

Then, the maximal F-measure on the curve is used as a summary statistic for the quality of the detector on the set of images. The current public version of the data set is divided in a training set of 200 images and a test set of 100 images. In order to ensure the integrity of the evaluation, only the images and segmentation results from the training set can be accessed during the optimization phase. In our case, these images have been employed to choose the parameters of the

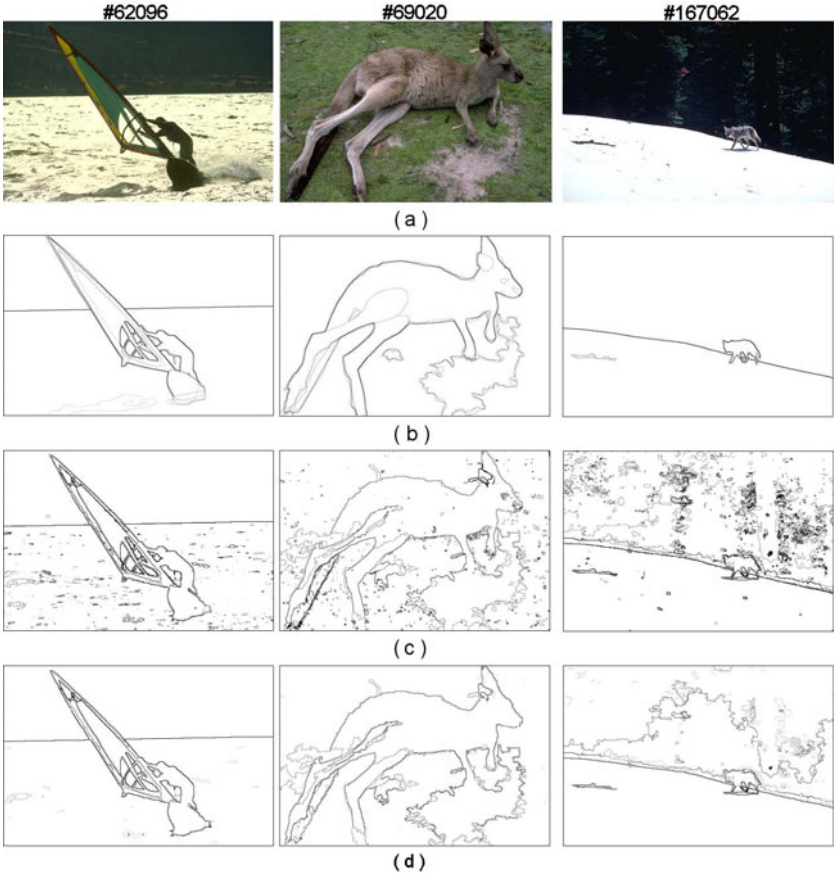


Fig. 2. a) Original images; b) boundaries of human segmentations; c) boundaries of pre-segmentation images; and d) boundaries of the regions obtained after the perceptual grouping

Table 1. Values of F for the images in Figure 2

	#62096	#69020	#167062
$NoPG$	0.85	0.63	0.41
PG	0.95	0.77	0.73

Table 2. Required time for each image in seconds

	#62096	#69020	#167062
$Pre-segmentation$	41.2	39.5	41.9
$Perceptual\ \ Grouping$	34.7	27.8	163.9
$Total\ \ time$	75.9	67.3	205.8

algorithm (i.e., the threshold U_p (see Section 2.1), the threshold U_s, α and β (see Section 2.2)). The optimal training parameters have been chosen. Fig. 2 shows the set of boundaries obtained in different segmentations of the original images as well as the ones marked by human subjects. It can be noted that the proposed approach is able to group perceptually important regions in spite of the large intensity variability presented on several areas of the input images. The pre-segmentation stage provides an over-segmentation of the image which overcomes the problem of noisy pixels [11], although bigger details are preserved in the final segmentation results.

The F-measure associated to each image in Fig. 2 can be seen in the Table 1. This Table shows the F-measure for the perceptual grouping stage (PG) as well as for the pre-segmentation stage ($NoPG$). These values of F reflect that adding a perceptual grouping stage improve significantly the results obtained in the segmentation.

On the other hand, Fig. 3 shows several images which have associated a low F-measure value. The main problems of the proposed approach are due to its inability to deal with textured regions which are defined at high natural scales. Thus, the tiger or the leopard in Fig. 3 are divided into a set of different regions. These regions do not usually appear in the human segmentations. The maximal F-measure obtained from the whole test set is 0.65. To improve it, other descriptors, such as the region area or shape, must be added to the distance $\psi^{\alpha,\beta}(\cdot,\cdot)$.

Regarding the execution times, the Table 2 summarize the processing time required for each of the images in Fig. 2. These times correspond to run the algorithm in a 1.60GHz Pentium PC, i.e. a sequential processor. It have to be noted that the proposed algorithm is mainly designed for parallel computing. Thus, it will run more efficiently in a parallel computer.

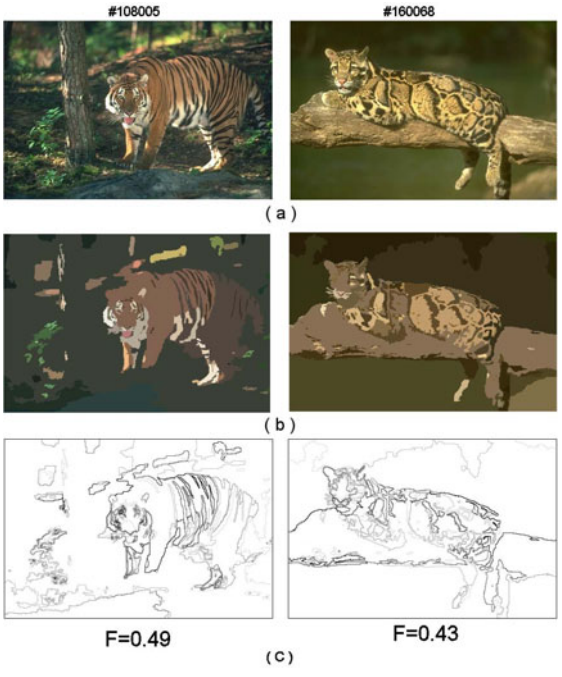


Fig. 3. a) Original images; b) example of segmented image; and c) set of obtained boundaries

4 Conclusions and Future Work

This paper presents a new perception-based segmentation approach which consists of two stages: a pre-segmentation stage and a perceptual grouping stage. In our proposal, both stages are conducted in the framework of a hierarchy of successively reduced combinatorial maps. The pre-segmentation is achieved using a color-based distance and it provides a mid-level representation which is more effective than the pixel-based representation of the original image. The combinatorial map which constitutes the top level of the hierarchy defined by the pre-segmentation stage is the first level of the hierarchy associated to the perceptual grouping stage. This second stage employs a distance which is also based on the colour difference between regions, but it includes information of the boundary of each region, and information provided by the Canny detector. Thus, this approach provides an efficient perceptual segmentation of the input image where the topological relationships among the regions are preserved.

Future work will be focused on adding other descriptors to the distance $\psi^{\alpha,\beta}(\cdot,\cdot)$, studying its repercussion in the efficiency of the method. Besides, it is necessary that the perceptual grouping stage also takes into account a texture measure defined at different natural scales to characterize the image pixels. This texture information could be locally estimated at the higher levels of the hierarchy.

Acknowledgments. This work has been partially granted by the Spanish Junta de Andalucía under project P07-TIC-03106 and by the Ministerio de Ciencia e Innovación (MICINN) and FEDER funds under projects no. TIN2008-06196 and AT2009-0026. This work extends the graph pyramid segmentation method proposed by Yll Haxhimusa, Adrian Ion and Walter Kropastch, Vienna University of Technology, Pattern Recognition and Image Processing Group, Austria [6].

References

1. Arbeláez, P.: Boundary extraction in natural images using ultrametric contour maps. In: Proc. 5th IEEE Workshop Perceptual Org. in Computer Vision, pp. 182–189 (2006)
2. Arbeláez, P., Cohen, L.: A metric approach to vector-valued image segmentation. Int. Journal of Computer Vision 69, 119–126 (2006)
3. Bister, M., Cornelis, J., Rosenfeld, A.: A critical view of pyramid segmentation algorithms. Pattern Recongition Letters 11(9), 605–617 (1990)
4. Borůvka, O.: O jistém problému minimálnim. Práce Mor. Přírodvěd. Spol. v Brně (Acta Societ. Scienc. Natur. Moravicae) 3(3), 37–58 (1926)
5. Brun, L., Kropatsch, W.: Introduction to combinatorial pyramids. In: Bertrand, G., Imiya, A., Klette, R. (eds.) Digital and Image Geometry. LNCS, vol. 2243, pp. 108–128. Springer, Heidelberg (2002)
6. Haxhimusa, Y., Ion, A., Kropatsch, W.G.: Evaluating hierarchical graph-based segmentation. In: Tang, Y.Y., et al. (eds.) Proceedings of 18th International Conference on Pattern Recognition (ICPR), Hong Kong, China, vol. 2, pp. 195–198. IEEE Computer Society, Los Alamitos (2006)
7. Haxhimusa, Y., Kropatsch, W.G.: Segmentation graph hierarchies. In: Fred, A., Caelli, T.M., Duin, R.P.W., Campilho, A.C., de Ridder, D. (eds.) SSPR&SPR 2004. LNCS, vol. 3138, pp. 343–351. Springer, Heidelberg (2004)
8. Huart, J., Bertolino, P.: Similarity-based and perception-based image segmentation. In: Proc. IEEE Int. Conf. on Image Processing, vol. 3, pp. 1148–1151 (2005)
9. Kropatsch, W.: Building irregular pyramids by dual graph contraction. IEEE Proc. Vision, Image and Signal Processing 142(6), 366–374 (1995)
10. Lau, H., Levine, M.: Finding a small number of regions in an image using low-level features. Pattern Recognition 35, 2323–2339 (2002)
11. Marfil, R., Bandera, A.: Comparison of perceptual grouping criteria within an integrated hierarchical framework. In: Torsello, A., Escolano, F., Brun, L. (eds.) GbRPR 2009. LNCS, vol. 5534, pp. 366–375. Springer, Heidelberg (2009)
12. Marfil, R., Molina-Tanco, L., Bandera, A., Rodríguez, J.A., Sandoval Hernández, F.: Pyramid segmentation algorithms revisited. Pattern Recognition 39(8), 1430–1451 (2006)
13. Martin, D., Fowlkes, C., Malik, J.: Learning to detect natural image boundaries using brightness, color, and texture cues. IEEE Trans. on Pattern Analysis Machine Intell. 26(1), 1–20 (2004)
14. Martin, D., Fowlkes, C., Tal, D., Malik, J.: A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In: Proc. Int. Conf. Computer Vision (2001)
15. Pham, T., Smeulders, A.: Learning spatial relations in object recognition. Pattern Recognition Letters 27, 1673–1684 (2006)
16. Zlatoff, N., Tellez, B., Baskurt, A.: Combining local belief from low-level primitives for perceptual grouping. Pattern Recognition 41, 1215–1229 (2008)