Topology-preserving perceptual segmentation using the Combinatorial Pyramid

Esther Antúnez, Rebeca Marfil and Antonio Bandera

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

Abstract Scene understanding and other high-level visual tasks usually rely on segmenting the captured images for dealing with a more efficient mid-level representation. Although this segmentation stage will consider topological constraints for the set of obtained regions (e.g., their internal connectivity), it is typical that the importance of preserving the topological relationships among regions will be not taken into account. Contrary to other similar approaches, this paper presents a bottom-up approach for perceptual segmentation of natural images which preserves the topology of the image. 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. 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 small regions and textured surfaces, experimental results reveal that the proposed perceptual grouping provides satisfactory scores.

Keyworks topology-preserving image segmentation, perceptual grouping, irregular pyramids, combinatorial pyramids.

1 Introduction

When the goal of the segmentation process is to divide the input image in a manner similar to human beings, image segmentation cannot be defined as the low-level process of grouping pixels into clusters which present homogeneous photometric properties. 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 [9]. The complexity of segmenting real objects from their background can be reduced if the particular application is taken into account. For instance, topology as a prior is available in many applications. Thus, the anatomy of human tissues (e.g. the cerebral cortex or the vasculature) provides important topological constraints which can help to ensure the correctness in medical image segmentation. Besides, when all the related structures are considered simultaneously, there could exist spatial relationships which can be also captured by topology (e.g., the brain is enclosed inside the skull, and the cerebellum and cerebrum are neighboring and separated organs, linked by the brainstem).

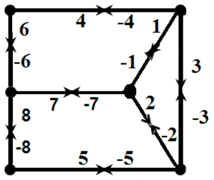
However, there exists different frameworks where it could be interesting to maintain the generality of use of the segmentation algorithm (e.g. an object-based attention mechanism or a visual-landmarks detector for autonomous robot navigation). In this regard, several authors have proposed generic segmentation methods which are based neither on a priori knowledge of the image content nor on any object model [1, 7]. These segmentations are called 'perceptual segmentations'.

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 [14].

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 [7, 11].

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 whose 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. It can be noted that this framework is closely related to the previous works of Arbeláez and Cohen [1, 2], Huart and Bertolino [7] and Marfil and Bandera [10]. 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 [7] and Marfil and Bandera employ the Bounded Irregular Pyramid (BIP) [10]. 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 [13], 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 [8]. 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. To avoid this problem, the described segmentation approach accomplishes the pre-segmentation and perceptual grouping stages by means of the Combinatorial Pyramid [4]. This irregular pyramid is defined by an initial combinatorial map which can be successively reduced using the scheme proposed by Kropatsch [8]. The combinatorial map encodes explicitly the orientation of edges around the graph vertices. Then, it uses a planar graph to represent each level of the pyramid instead of a pair of dual graphs. This 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



(a)

dart	1	-1	2	-2	3	-3	4	-4	5	-5	6	-6	7	-7	8	-8
α	-1	1	-2	2	-3	3	-4	4	-5	5	-6	6	-7	7	-8	8
σ	3	-7	-1	-5	-4	-2	6	1	-8	-3	4	8	-6	2	7	5

(b)

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

proposed method are presented in Section 3. Finally, the paper concludes along with discussions and future work in Section 4.

2 Perceptual segmentation

Irregular pyramids represent the image as a stack of graphs with decreasing number of vertices. Such hierarchies present many interesting properties within the Image Processing and Analysis framework such as: reducing the influence of noise by eliminating less important details in upper levels of the hierarchy, making the processing independent of the resolution of the regions of interest in the image, converting global features to local ones, reducing computational costs, etc. The construction of the pyramid follows the philosophy of reducing the amount of data between consecutive levels of the hierarchy by a reduction factor $\lambda > 1$. In this hierarchy, every vertex v_k in level k is linked with a set of vertices on the underneath level $k-1$, $\{v_i\}_{k-1}$. Those vertices $\{v_i\}_{k-1}$ are called the children of v_k, which will be then called their parent. The highest level of the pyramid is called its apex. Each pyramid level is encoded as a graph. As it has been aforementioned, combinatorial maps can be used to perform this encoding.

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. That is, a n-dimensional combinatorial map is 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 and the other β_i are involutions on D. An involution is a permutation whose cycle has the length of two or less. A combinatorial pyramid is a hierarchical stack of combinatorial maps successively reduced by a sequence of contraction or removal operations [4, 8]. Combinatorial pyramids combine the advantages of dual graph pyramids with an explicit orientation of the boundary segments of the embedded object thanks to one of the permutations which defines the map [4]. Moreover, using combinatorial maps, the dual graph is both implicitly encoded and updated.

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

2.1 Pre-segmentation

Two-dimensional (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:

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

Figure 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 $\sigma^*(1) = \{1, 3, -4\}$.

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, $\phi^*(3) = \{3, -2, -1\}$.

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. These faces have an attribute that store 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 [6, 5]. However, in this proposal, 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 of the graph is compared with the threshold U_p and if its value is smaller, that edge if added to a removal kernel. In a second step, hanging edges are removed. Finally, a contraction kernel is applied to remove parallel edges, obtaining the new level of the pyramid. This process is iteratively repeated until no more removal/contraction operation is possible. Algorithm 1 shows how to build the combinatorial pyramid. The hierarchy of graphs is built based on a spanning tree of the initial graph obtained using the algortihm of Borůvka [3]. Building the spanning tree allows to find the region borders quickly and effortlessly based on local differences in a color space. This process 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

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.

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

Algorithm 1 Construction of the combinatorial pyramid

```
$k = 0$, **Input**: Attributed combinatorial map $G_0$
**repeat**
    **for all** faces $f \in F_k = \varphi_k^*(D_k)$ **do**
        $E_{min}(f) = argmin\{attr(e)|e = (d, -d) \in E_k = \alpha_k^*(D)\ and\ f = \varphi_k^*(d)\}$ {Borůvka's algorithm [3]}
    **end for**
    **for all** $e = (d, -d) \in E_{min}$ with $attr(e) < U_p$ **do**
        include $d$ and $-d$ in a removal kernel $RK1_{k,k+1}$
    **end for**
    reduce combinatorial map $G_k$ with removal kernel: $G'_{k+1} = R[G_k, RK1_{k,k+1}]$
    **for all** hanging edge $e = (d, -d)$ in $G'$ **do**
        include $d$ and $-d$ in a removal kernel $RK2_{k,k+1}$
    **end for**
    reduce combinatorial map $G'_{k+1}$ with removal kernel: $G''_{k+1} = R[G'_{k+1}, RK2_{k,k+1}]$
    **for all** parallel edge $e = (d, -d)$ in $G''$ **do**
        include $d$ and $-d$ in a contraction kernel $CK1_{k,k+1}$
    **end for**
    reduce combinatorial map $G''_{k+1}$ with contraction kernel: $G_{k+1} = C[G''_{k+1}, CK1_{k,k+1}]$
    **for all** $e_{k+1} \in E_{k+1} = \alpha_{k+1}^*(D_{k+1})$ **do**
        set edge attributes after contraction: $attr(e_{k+1}) = colorDistance(face(d_{k+1}), face(-d_{k+1}))$
    **end for**
    $k = k + 1$
**until** $G_k = G_{k-1}$
```

complemented with internal region properties and with attributes of the boundary shared by both regions. The distance between two regions (faces) $\mathbf{f}_i \in G_l$ and $\mathbf{f}_j \in G_l$, $\psi^{\alpha,\beta}(\mathbf{f}_i, \mathbf{f}_j)$, is defined as

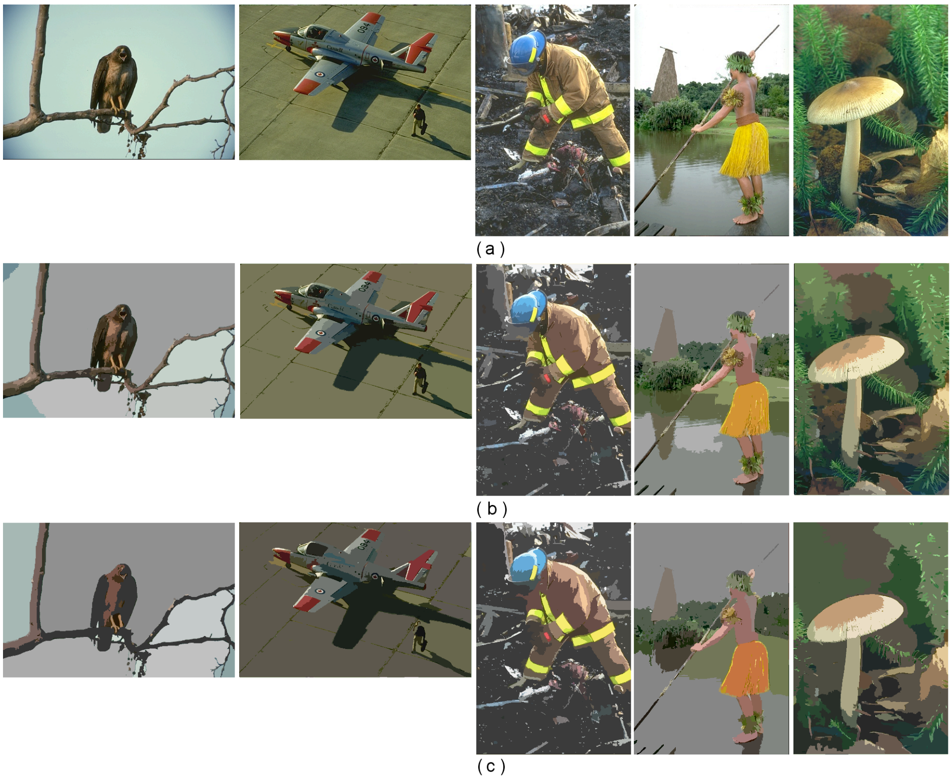
$$\psi^{\alpha,\beta}(\mathbf{f}_i, \mathbf{f}_j) = \frac{d(\mathbf{f}_i, \mathbf{f}_j) \cdot b_{\mathbf{f}_i}}{\alpha \cdot (c_{\mathbf{f}_i\mathbf{f}_j}) + (\beta \cdot (b_{\mathbf{f}_i\mathbf{f}_j} - c_{\mathbf{f}_i\mathbf{f}_j}))} \tag{2}$$

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 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] [12]. 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* [11]. This technique considers two quality measures:

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



(a)

(b)

(c)

Figure 2: a) Original images; b) pre-segmentation images; and c) obtained regions after the perceptual grouping.

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 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 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 partitions on the higher level of the hierarchy for five different images. 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 [10], although bigger details are preserved in the final segmentation results.

The F-measure associated to the individual results ranged from bad to significantly good values. Thus, the F-measure value of all images in Fig. 2 is over 0.8. On the contrary, 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 snake, zebras or the backgroud in Fig. 3 are divided into a set of different regions. Besides, the approach preserves the existence of regions of very small area at higher levels of the hierarchy (this can be seen in the bottom left part of the third image in Fig. 3). 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 $\psi^{\alpha,\beta}(\cdot,\cdot)$.

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 [5].

References

[1] P. Arbeláez. Boundary extraction in natural images using ultrametric contour maps. In *Proc. 5th IEEE Workshop Perceptual Org. in Computer Vision*, pages 182–189, 2006.

[2] P. Arbeláez and L. Cohen. A metric approach to vector-valued image segmentation. *Int. Journal of Computer Vision*, 69:119–126, 2006.

[3] O. Borůvka. O jistém problému minimálnim. *Práce Mor. Přírodvěd. Spol. v Brně (Acta Societ. Scienc. Natur. Moravicae)*, 3(3):37–58, 1926.

[4] L. Brun and W. Kropatsch. Introduction to combinatorial pyramids. *Lecture Notes in Computer Science*, 2243:108–128, 2001.

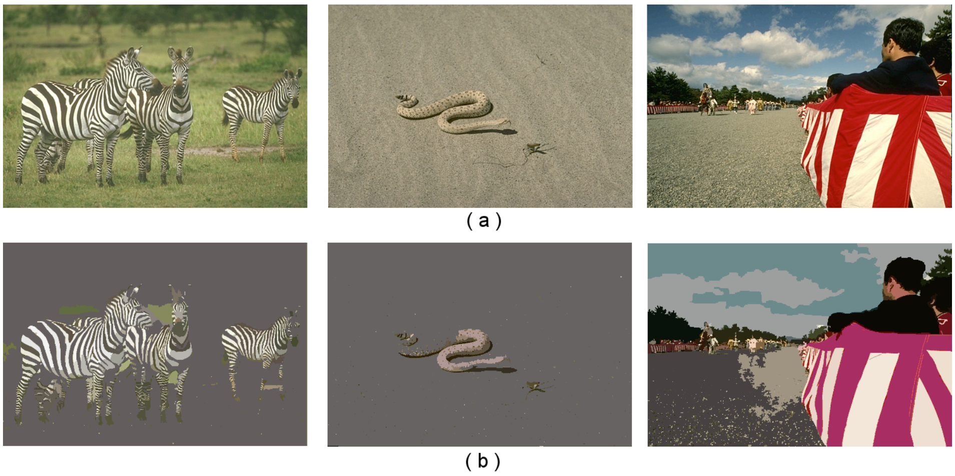


Figure 3: a) Original images; and b) obtained regions after the perceptual grouping.

[5] Y. Haxhimusa, A. Ion, and W. G. Kropatsch. Evaluating hierarchical graph-based segmentation. In Y. Y Tang et al, editor, *Proceedings of 18th International Conference on Pattern Recognition (ICPR)*, volume 2, pages 195–198, Hong Kong, China, 2006. IEEE Computer Society.

[6] Y. Haxhimusa and W.G. Kropatsch. Segmentation graph hierarchies. In A. L. N. Fred et al., editor, *Joint IAPR Int. Workshops, SSPR2004 and SPR2004*, volume 3138 of *LNCS*, pages 343–351. Springer, 2004.

[7] J. Huart and P Bertolino. Similarity-based and perception-based image segmentation. In *Proc. IEEE Int. Conf. on Image Processing*, volume 3, pages 1148–1151, 2005.

[8] W. Kropatsch. Building irregular pyramids by dual graph contraction. *IEEE Proc. Vision, Image and Signal Processing*, 142(6):366–374, 1995.

[9] H. Lau and M. Levine. Finding a small number of regions in an image using low-level features. *Pattern Recognition*, 35:2323–2339, 2002.

[10] R. Marfil and A. Bandera. Comparison of perceptual grouping criteria within an integrated hierarchical framework. *Lecture Notes in Computer Science*, 5534:366–375, 2009.

[11] D. Martin, C. Fowlkes, and J. Malik. Learning to detect natural image boundaries using brightness, color, and texture cues. *IEEE Trans. on Pattern Analysis Machine Intell.*, 26(1):1–20, 2004.

[12] D. Martin, C. Fowlkes, D. Tal, and J. Malik. 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.

[13] T. Pham and A. Smeulders. Learning spatial relations in object recognition. *Pattern Recognition Letters*, 27:1673–1684, 2006.

[14] N. Zlatoff, B. Tellez, and A. Baskurt. Combining local belief from low-level primitives for perceptual grouping. *Pattern Recognition*, 41:1215–1229, 2008.