

# EXTRACTION OF SEMANTIC OBJECTS FROM STILL IMAGES

*Alvaro Pardo\**

IMERL & IIE  
Facultad de Ingeniería  
Universidad de la República  
C.C. 30, Montevideo, Uruguay

## ABSTRACT

In this work, we study the extraction of semantic objects from still images. We combine different ideas to extract them in a structured manner together with a perceptual metric that ranks them according with its perceptual relevance.

The algorithm has four steps, the regularization of the initial segmentation using probability diffusion [1], simplification of the segmentation via region merging, computation of the perceptual metric based on [2] and construction of the structure that represents the image (the binary partition tree [3]).

## 1. INTRODUCTION

The extraction of semantic objects from images or sequences is one of the most challenging problems in image analysis. These systems are of key importance for the new content-based applications like: object-based image and video compression (standards JPEG2000 and MPEG-4), multimedia applications that permit some kind of object manipulation (video indexing in the context of MPEG-7), and video segmentation for tracking and surveillance.

The main difficulty of this problem resides in the fact that simple features do not uniquely determine semantic objects. For instance, segmentation into homogeneous regions does not lead to semantic object extraction. Humans correctly detect semantic objects in a wide range of conditions; lightning conditions, occlusion, change in colour within the same entity, etc. For an algorithm, this task is more difficult and it has been one of the most studied problems in image analysis. For this reason, several works assumed that some kind of user interaction must be added to extract semantic objects. This has raised the classification of algorithms into automatic or semi-automatic. Although the latter's do not automatically accomplish the objective, they provide a good initial condition. For instance, in several works on extraction of semantic video objects, the first image is segmented in a semi-automatic way and then the objects are tracked in the sequence.

Now, we briefly review some of the works on semantic object extraction from still images<sup>1</sup>. The most classical approaches are based on region merging [4], region growing algorithms and spatial segmentation with multiple thresholds [5]. In [6, 7] the basic tools for the extraction of regions are connected operators which filter the image merging the flat zones. In this way, they do not introduce new contours and therefore contours are preserved. Also other morphological transformations such as region-growing watershed, geodesic skeleton and propagation of markers have been used [8, 6, 9]. The previous approaches use simple features: colour, edge information, texture, and motion for video segmentation.

For several region-based applications, we not only need the segmentation, but also a good data structure to represent it. In [3] Salembier et al. introduce the idea of binary partition trees (BPT) as a suitable data structure.

Regarding the perceptual metric, we cite [2]. For every region in the segmented image, simple features are computed and combined to obtain a map of perceptual importance. Although, there are many works dealing with this problem, this approach is simple and is the one we are going to use.

### 1.1. Our Approach

In this work, we combine different ideas to extract semantic objects from images in a structured manner.

From a fine initial segmentation, we first apply a regularization step using our previous work on probability diffusion applied to classification problems [1]. On one hand, the number of regions decreases when regularizing the segmentation (small noisy regions are merged). On the other hand, in this step new regions can be added. We must say that the first one is the main goal since it reduces the complexity of the succeeding steps.

With the resulting segmentation after the regularization step, we construct the region adjacency graph (RAG), where each node represents a region and links connect neighbouring regions. We model each region with its mean and a

---

\* (apardo@iie.edu.uy) Supported by CSIC - Universidad de la República and Fondo Clemente Estable 2000 No. 6034- Conicyt

---

<sup>1</sup>For video sequences, motion information plays an important role since semantic objects usually move together.

region-merging algorithm is applied to obtain the desired number of regions. At this point, the user should interact with the algorithm to stop it when the resulting segmentation is the desired one.

The segmentation result of the merging algorithm contains regions nearly matching the semantic objects in the image. With this segmentation, we compute a perceptual metric (PM) that ranks objects according with its perceptual importance. We improved Osberger's work [2].

Once we have the semantic segmentation, we construct a BPT with it [3]. We again apply a merging algorithm using the PM. Each time two regions are merged a new node of the tree is added. This procedure is applied until there is only one region. Thus, in the leaves of the tree we have the initial regions and in the remaining nodes their unions. After computing the BPT, the user can improve the segmentation manually (BPT make it very simple and fast [3]).

Although any initial segmentation can be used, we use the segmentation obtained with algorithms we presented in [10]. The advantage of it is that regions already match the objects of interest.

## 2. SEGMENTATION REGULARIZATION

In the first step of the algorithm we add coherence to the initial segmentation using the vector probability diffusion scheme (VPD) [1] by adding spatial coherence to the posteriors probabilities of classes present in the image.

We say that a given region from the initial segmentation  $\{R_1, \dots, R_n\}$ , is a valid class if its area is bigger than a given threshold. Each class  $c_i \in \mathcal{C} = \{c_i : i = 1, \dots, m\}$  is represented with the mean of its members:  $\mu_i$ . For every pixel  $x$  we have a probability vector  $p(x) \in \mathcal{P} = \{p \in \mathbb{R}^m : \|p\|_1 = 1, p_i \geq 0\}$  where  $p_i(x)$  equals the probability of pixel  $x$  to belong to the class  $c_i$ :

$$p_i(x) = \frac{1}{|I(x) - \mu_i|} \left( \sum_{j=1}^m \frac{1}{|I(x) - \mu_j|} \right)^{-1} \quad (1)$$

To add spatial coherence into the classification process VPD diffuses the distance between points in  $\mathcal{P}$  with the following diffusion equations:

$$\frac{\partial p_i}{\partial t} = \nabla \cdot \left( \frac{\nabla p_i}{\sqrt{\sum_{i=1}^m \|\nabla p_i\|^2}} \right) \quad i = 1, \dots, m.$$

For further details and implementation see [1].

## 3. REGION MERGING ALGORITHM

To apply the merging algorithm we need to define the region model and the merging criterion, which depends on a distance between regions. The region model  $\mu_i$  is defined as

the mean of the pixels in the region  $R_i$ . When two regions are merged, the new model must be computed. To make the model estimation robust, the new model equals the one of the bigger region [11, 3].

The merging criterion minimizes the cost of each merging. That is, in each step we minimize the cost function (2) merging the pair of nodes with minimum cost.

$$C(R_1, R_2) = P(R_1)D(R_1, R_1 \cup R_2) + P(R_2)D(R_2, R_1 \cup R_2) \quad (2)$$

$D$  is the distance between regions:  $D(R_i, R_j) = |\mu_i - \mu_j|^2$ , and  $P(R_i)$  is the probability of region  $R_i$ :  $P(R_i) = \text{Area}(R_i)/\text{Area}(\Omega)$ . This cost function measures the error between the given partition and the new model. For highly textured regions the mean alone will not be enough to discriminate between regions, we should include a measure of texture like the variance within the region.

## 4. PERCEPTUAL METRIC

In this section, we present a PM to automatically determine the perceptual importance of different regions in the image. This metric is based on Osberger's work [2] and uses several features that influence human visual attention. For each region in the image, a set of features is computed and then combined to obtain the importance map that ranks each region with respect to its perceptual importance.

Firstly, to apply this idea we need a segmentation of the image. This point is crucial; regions in the segmentation should represent semantic regions or part of them, otherwise, the perceptual metric will not correlate with our perception. For this reason, we do not use the initial segmentation to compute the importance map as it contains too many small regions. Instead, we compute the perceptual metric using a coarser segmented image, the one obtained after some steps of the merging algorithm.

### 4.1. Factors which influence our attention

These factors can be classified into: low level and high level. Among low-level factors, we have: contrast, size, shape, and colour. High-level factors are of course more difficult to model. For instance, the presence of people in the image is a strong factor; our attention is drawn to their eyes, mouth, and hands. In our case, we use location and the distinction of foreground and background as high-level factors.

**Contrast:** Region contrast is a very strong factor; regions with high contrast with their neighbour regions attract our attention, and therefore they might belong to regions of perceptual importance. The contrast of a region  $R_i$  which has a set of neighbours  $\{R_{i_1}, \dots, R_{i_N}\}$  is computed as:

$$\text{Contrast}(R_i) = \frac{1}{N} \sum_{j=1}^N \alpha_j |\mu_i - \mu_{i_j}|$$

$$\alpha_j = \text{Length}(\partial R_i \cap \partial R_{i_j}) / \text{Perimeter}(R_i)$$

where  $\mu_{i_j}$  are the means of the regions  $R_{i_j}$  and  $\alpha_j$  weights the contribution of each neighbouring region to the contrast measure. That is, the more contact between regions the more it should contribute to the contrast measure.

Osberger measures the contrast as the difference of the mean of a region and the mean of the neighbour regions. This is not a robust measure since the mean of the neighbouring regions is strongly affected by an “outlier” region. For example, take a region with mean 128 and two neighbours with means 255 and 0. In this case the mean of the neighbouring regions equals the mean of the region and therefore according to Osberger metric the contrast will be 0. Obviously, this does not reflect what we perceive. With our definition of contrast (neglecting the factors  $\alpha_j$ ), the contrast is 128.

**Size:** It has been found that region size is an important factor. Large regions are more likely to attract our attention than the small ones. The size measure is computed as:

$$\text{Size}(R_i) = \max \{ \text{Area}(R_i) / A_{max}, 1 \}$$

where  $A_{max}$  is set to the 1% of the total area and is used to prevent excessive weighting to very large regions.

**Shape:** It has been argued that long and thin regions are visual attractors [12], but also that our perception tends to favour compact regions [13]. Osberger applies the first idea and uses the shape factor:  $\text{Perimeter}(R_i)^{1.75} / \text{Area}(R_i)$  trying to capture long and thin regions. Conversely, we apply the second idea using the isoperimetric ratio of the region which scores compact regions as more important. According to our experiments, this selection performs better. Note that the isoperimetric ratio is nearly the inverse of Osberger’s measure.

$$\text{Shape}(R_i) = \text{Area}(R_i) / \text{Perimeter}(R_i)^2$$

**Foreground/Background:** Typically, objects in the foreground attract our attention. To determine if a region is part of the background we measure the number of pixels of the region border that belong to the image border. In this way the foreground/background measure is computed as:

$$\text{FB}(R_i) = 1 - \min \left\{ \frac{\text{Length}(\partial R_i \cap \partial \Omega)}{0.5 * \text{Perimeter}(\Omega)}, 1 \right\}$$

**Location:** Different experiments have shown that typically viewers focus at the centre of the image. To compute this factor we measure the number of pixels of the region which are within the 25% centre of the image:  $\text{Centre}(R_i)$ . Regions that are entirely in the centre of the image will have the maximum weight.

$$\text{Location}(R_i) = \text{Centre}(R_i) / \text{Area}(R_i)$$

**Importance Map:** After normalizing each of the factors presented above to the range  $[0, 1]$  the importance map is computed as the sum of their squared values. This assigns higher scores to regions with high scores in some factors.

## 5. ALGORITHM

In this section, we describe the whole algorithm. We argue that semantic objects must be used to construct the BPT. We want a small tree capturing the semantic objects in the image. Therefore, we compute the BPT after some steps of the region-merging algorithm.

1. Given the initial partition we apply the VPD to add coherence to the initial segmentation and recompute the partition.
2. Apply the merging algorithm until we obtain the desired number of regions,  $R_{th_1}$ . In this step, the user interaction may want to control the threshold to stop the merging when the segmentation captures the semantic objects in the image.
3. Compute the perceptual metric.
4. A new merging step is applied but now the metric is weighted with the perceptual metric:

$$\begin{aligned} \bar{C}(R_1, R_2) &= PM(R_1)P(R_1)D(R_1, R_1 \cup R_2) \\ &+ PM(R_2)P(R_2)D(R_2, R_1 \cup R_2) \end{aligned}$$

Apply merging algorithm until the number of regions is  $R_{th_2}$  or until a single region is found. In the later case, the merging order is used to build the BPT. When two regions are merged, the perceptual metric is updated with:

$$PM(R_1 \cup R_2) = \min \{ PM(R_1), PM(R_2) \}$$

## 6. RESULTS

In figure 1 we show the results for Claire image. The initial segmentation was obtained with the algorithm presented in [10]; it contains all level sets which: contain T-junctions at their boundaries, and have good contrast and shape. We show the initial segmentation, the result after VPD (the number of regions is reduced while the most important regions are kept), and finally the result of the merging algorithm (at this point the user controls the merging in order to obtain regions matching the semantic objects). For this simple image, we achieved our goal of segmenting the image into semantic objects. For this segmentation, we computed the PM and at last the BPT. We also show the final segmentation for Carphone image.

## 7. CONCLUSIONS

Although the algorithms used in this work are not completely new, we integrated them to extract semantic objects in previously segmented images. We improved the computation of the perceptual metric proposed by Osberger. We

used the perceptual metric in the construction of the BPT and we stated that this tree must be based on important regions (the ones obtained after the supervised region merging). As we showed, the inclusion of the perceptual metric in the BPT construction helps on moving close to the root the perceptually important objects.

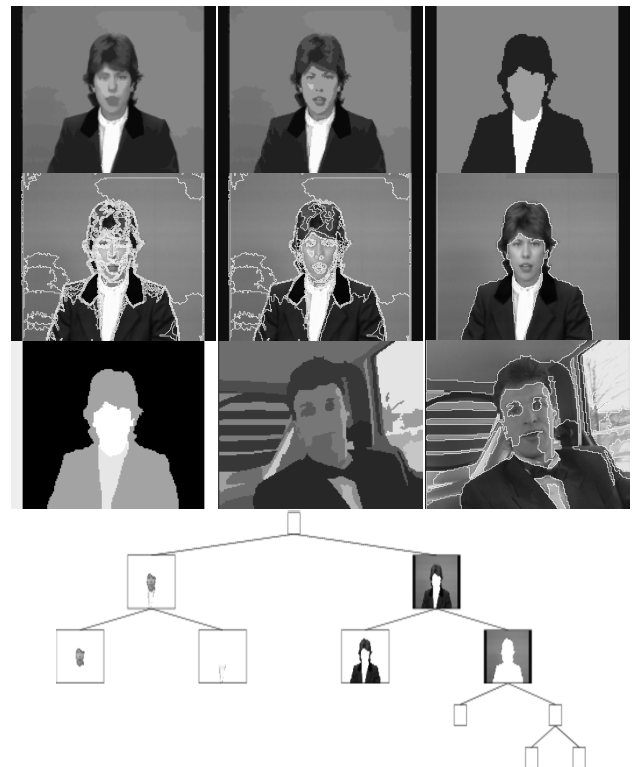
For simple images, like Claire, the algorithm easily extracts the semantic objects. On the other hand, for complex image, like Carphone, the algorithms proposed do not achieve the objective completely. However, the structured results here presented would aid the user to correct the segmentation and properly extract the semantic objects.

## 8. REFERENCES

- [1] A. Pardo and G. Sapiro, "Vector Probability Diffusion," *IEEE Signal Processing Letters*, vol. 8, no. 4, pp. 106–109, Apr. 2001.
- [2] W. Osberger and A. J. Maeder, "Automatic Identification of Perceptually Important Regions in an Image," in *ICPR*, 1998, vol. 1, pp. 701–704.
- [3] P. Salembier and L. Garrido, "Binary Partition Tree as an Efficient Representation for Image Processing, Segmentation, and Information Retrieval," *IEEE Trans. Image Processing*, vol. 9, no. 4, pp. 561–576, Apr. 2000.
- [4] H. Sang Park and J. Beom Ra, "Homogenous Region Merging Approach for Image Segmentation Preserving Semantic Object Contours," in *Proc. VLBV'98*, 1998, pp. 149–152.
- [5] F. Long, D. Feng, H. Peng, and W. Siu, "Extracting Semantic Video Objects," *IEEE Comp. Graphics and Applications*, vol. 21, no. 1, pp. 48–55, Jan./Feb. 2001.
- [6] P. Salembier, P. Brigger, J. Casas, and M. Pardas, "Morphological Operators for Image and Video Compression," *IEEE Trans. Image Processing*, vol. 5, no. 6, pp. 881–898, Jun. 1996.
- [7] P. Salembier, A. Oliveras, and L. Garrido, "Antiextensive Connected Operators for Image and Sequence Processing," *IEEE Trans. Image Processing*, vol. 7, no. 4, pp. 555–570, Apr. 1998.
- [8] P. Salembier, L. Torres, F. Meyer, and C. Gu, "Region-Based Video Coding Using Mathematical Morphology," *Proc. IEEE*, vol. 83, no. 6, pp. 843–857, Jun. 1995.
- [9] D. Gatica-Perez, C. Gu, and M. Sun, "Semantic Video Object Extraction Using Four-Band Watershed and Partition Lattice Operators," *IEEE Trans. Circuits*

*Syst. Video Technol.*, vol. 11, no. 5, pp. 603–618, May 2001.

- [10] A. Pardo, "Semantic Image Segmentation using Morphological Tools," in *Proc. ICIP'2002*, 2002.
- [11] L. Garrido, P. Salembier, and D. Garcia, "Extensive operators in partition lattices for image sequence analysis," *Signal Processing*, vol. 66, no. 2, pp. 157–180, Apr. 1998.
- [12] J.W. Senders, "Distribution of attention in static and dynamic scenes," in *SPIE*, Feb. 1997, number 3016, pp. 186–194.
- [13] W. Kohler, *Gestalt Psychology*, Liveright Publishing Corporation, 1947.



**Fig. 1.** First two rows: Claire with 1994 regions, after two iterations of VPD (area threshold is 100 pixels) with 766 regions, and the image with 6 regions matching the semantic objects. We show the images with regions represented by its mean and boundaries. Third row: Claire PM (Bright indicates important. The results match our perception; the face is the most important region) and Carphone segmentation results (30 regions). Finally, we show Claire BPT.