BPT Enhancement Based on Syntactic and Semantic Criteria [⋆]

C.Ferran, X.Giró, F.Marqués and J.R.Casas

Signal Theory and Communications Department,
Technical University of Catalonia (UPC)
Campus Nord, C/ Jordi Girona, 1-3,
08034, Barcelona

Abstract. This paper presents two enhancements for the creation and analysis of Binary Partition Trees (BPTs). Firstly, the classic creation of BPT based on colour is expanded to include syntactic criteria derived from human perception. Secondly, a method to include semantic information in the BPT analysis is shown thanks to the definition of the BPT Semantic Neighborhood and the introduction of Semantic Trees. Both techniques aim at bridging the semantic gap between signal and semantics following a bottom-up and a top-down approach, respectively.

1 Introduction

1.1 Problem statement: The semantic gap

Scene analysis for object extraction is a challenging and unsolved problem. The main difficulty of this task is to connect semantic entities to low-level features or vice-versa; that is, to fill the so-called **semantic gap**. There mainly exist two conceptually opposed approaches: **Top-down** methods, which are knowledge-based and search for pre-defined models into the image given some prior information. **Bottom-up** methods, which are generic methods aiming at linking visual features to perceptual meaningful primitives.

Moreover, methods based only on either top-down or bottom-up approaches might fail in a general context. On one hand, bottom-up methods cannot infer object entities without prior models. On the other hand, top-down techniques may benefit from a richer pre-analysis based on perceptual considerations. Thus, both approaches are not mutually exclusive, on the contrary, they might be coupled in order to succeed in the object extraction task. For this reason, a **common framework** is required.

As presented in [1], a **Binary-Partition-Tree** (BPT) is a structured representation of a set of hierarchical partitions usually obtained by means of an iterative segmentation procedure based on the optimization of an initial partition. The BPT is built up by iteratively merging the most similar pair of regions

[⋆] This work has been partly supported by the EU Network of Excellence MUSCLE, and by the grant TEC2004-01914 of the Spanish Government.

given a homogeneity criterion. In this work the initial partition is assumed to include all the contours of the desired object. Therefore, the BPT representation can theoretically lead to a tree where each desired image object is represented by a single node at a certain level of the hierarchy. However, as pointed out in [2], from the bottom-up point of view there are some drawbacks related to the *simplicity* of the homogeneity properties used to construct the tree. This limitation can result in a tree where the desired object is not represented by a node. From the top-down point of view, not all the regions defined by the BPT are significant. Still the BPT representation offers the support for the extraction of objects and a common framework where regions and semantic entities can be linked. The goal of this paper is to overcome some of the previous limitations to enhance the BPT framework with the long term aim of bridging the semantic gap.

1.2 Proposal: Enhanced BPT as common framework

This paper presents the BPT as a common framework for bottom-up and top-down analysis. The framework has been improved from both points of view:

- **Bottom-up BPT construction:** Introducing and combining multiple and generic homogeneity criteria based on low- and middle-level features. Such features are refereed to as **syntactic** features, since they are defined by the relative positions of the regions they represent.
- **Top-down BPT analysis:** The problem of detections of a single instance of the same object is assessed. To do so, a model for semantic classes and its application on BPTs is presented.

The paper is organized as follows. Firstly, we present the motivations for both, the syntactic approach for BPT construction and the subsequent semantic BPT analysis. Secondly, we describe the enhanced BPT common framework and, some object detection and extraction examples are presented. Finally, the conclusions and the future worklines are drawn.

1.3 Motivations for the syntactic approach

When analysing an image, colour, motion or even depth information are usually prioritized disregarding other natural features of the objects belonging to the scene, such as symmetry or partial inclusion. Such features might be classified as being in the middle level vision problem.

This work relies on the assumption that, such complex features can be used to group image elements in order to form parts of objects or even complete objects. This assumption is based on the Gestalt psychology and on perceptual grouping approaches, as presented in [3] and [4], respectively.

In this context we assume that a **composite** object is an object which is built up of a set of distinct parts in various ways of **complex** composition. A well known technique for the study of such composition properties is syntactic image analysis [5]. Such work deals with the notions of primitive, grammar and syntax analysis from a formal point of view, ideas that have inspired the creation of the proposed syntactic framework for image segmentation.

This proposal is based on two key aspects. Firstly, the definition and extraction of the syntactic features expressed as rules, and secondly, the critical decision of how to combine the different features. Regarding this second aspect, there may be situations in which the correct combination of different features (for instance, "similar colour as" and "partially included into") to create an object is not easy even for a human observer. In this work we attempt to provide a solution for such cases. Our proposal aims to apply the most significant rule given a set of features. This solution is based on the statistical analysis of the features over the whole image.

As a conclusion, we can define the syntactic image segmentation as an iterative region merging procedure based on the assessment of multiple homogeneity criteria expressed through rules which are combined according to their statistics. At each iteration the possible conflict between the defined rules is solved resulting in the merging of the most significant pair of regions.

1.4 Motivations for the semantic approach

Many analysis tasks focus on the extraction of semantic information from visual content. In applications such as object detection or content-based image retrieval, the final goal is to extract semantic entities from the visual data, as this is the most natural form for humans to access content. In most of the cases, the result of a segmentation and its organization in BPTs lead to many more regions than semantic entities are present in the content. Most of the regions represented by the BPT nodes lack a semantic meaning, although its organization in BPTs facilitates their analysis.

Partitions usually present oversegmentation problems, splitting a single object among more than a region. Nevertheless, this situation may be solved thanks to the hierarchical structure of the BPT. However, the BPT may contain a few nodes related to the same object which, in turn, may present similar perceptual properties and, for this reason, they may be considered as multiple instances of the same class if the analysis is based only on low-level features. Fig. 1 depicts an example of these cases, which represents how a fading rectangle is split in two during the segmentation process.

Selecting which of these similar regions represents the instance of a semantic class requires a previous semantic knowledge. This paper will propose a method to model semantic classes and how these models can help in a better semantic analysis of BPTs.

2 Enhanced BPT Framework

A BPT is created from an initial partition of an image, plus a set of hierarchichal partitions that are "‘above" the initial partition. The finest level of detail is given

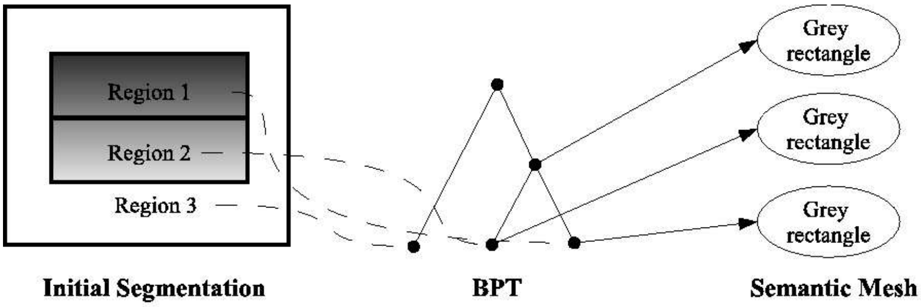

Fig. 1. Oversegmentation and BPT generates multiple instances of the same class.

by the initial partition (the leaves of the BPT). The nodes above the leaves are associated to regions resulting from the merging of two children regions until we reach the root node, which represents the entire image support. An example illustrating the BPT for an image made of two rectangles over a white background can be seen on Fig.1.

2.1 Bottom-Up: Syntactic BPT creation

The framework we present for BPT creation is an extension of the segmentation algorithm presented in [6], which performs an iterative region-merging. In this framework, an initial over-segmented partition is optimized by merging pairs of regions following a pre-defined optimization criterion. The BPT is built up by tracking the successive mergings of the algorithm. The aim is to iteratively estimate the region of support of homogeneous elements based on the features of the elements. The optimization criteria are based on the homogeneity of the elements and are defined as homogeneity criteria. For this purpose we define two type of descriptors:

1. Visual descriptors (VDs). These descriptors are computed locally, in a region neighboorhood, and provide individual dissimilarity measures for:
   Simple homogeneity criteria evaluated over pixel values within a single element and computed with *simple* operations between elements (pixels or regions). Colour, or region size are examples of descriptors computed using simple homogeneity criteria.
   Complex homogeneity criteria evaluated for two or more elements and computed using *complex* operations between elements (regions). Syntactic descriptors, such as symmetry or partial inclusion, are based on complex homogeneity criteria.
2. Statistical descriptors modelling the statistical distribution of the VDs and computed over the whole image, globally. For example, the entropy H_r of each rule r provides a measure of its significance in the current iteration. The statistical information allows to manage multiple homogeneity criteria by the use of global information.

2.2 Syntactic Segmentation

The major segmentation steps are simplification, feature extraction and decision, as presented in the scheme of Fig. 2.

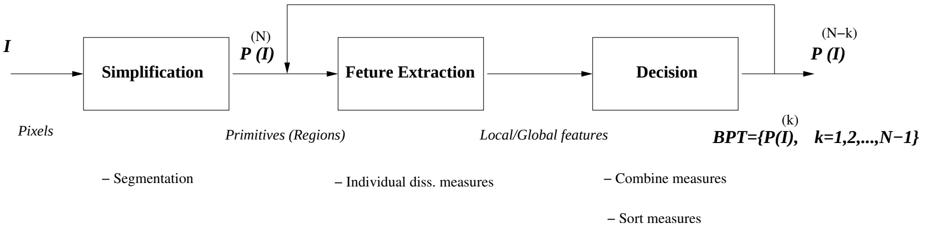

Fig. 2. Scheme of the major segmentation steps.

Simplification The purpose of the simplification step is to control the amount and nature of the information preserved for further analysis. Starting at the pixel level it generates a set of regions as primitives for the following steps. The resulting partition provides a robust information-rich set of region primitives overcoming the limitations of point-based primitives for higher level analysis.

Feature extraction The feature space is created by associating a dissimilarity measure to each pair of regions given one of the following rules: R_i **has similar mean colour than** R_j, R_i **has similar size than** R_j and **is** R_i **partially included into** R_j. Moreover we impose the constraint that only neighbouring regions can be merged so that the result of any merging is still a partition of connected regions. The dissimilarity measure between regions R_i and R_j in the $k-th$ iteration using the $c-th$ rule is noted as $d_c^k(i,j)$. Its value tends to zero if the proposition is found to be true.

Even though the rules are computed locally over a pair of regions, an estimate of the distribution of the dissimilarity measures for each rule is computed by means of the histogram. In the case of colour homogeneity, the histogram counts the number of times that the same colour difference occurs among the whole set of regions. The distributions for size and partial inclusion are similarly estimated, extracting features which are both local and global as a result.

Decision The decision step selects the most significant rules given the current state of the feature space (see Sec. 2.3), deciding how to combine the extracted dissimilarity features in a unique dissimilarity measure $(D^k(i,j))$ for each pair of connected regions. Finally, the most similar pair of regions is selected for merging.

As presented in Fig. 2, feature extraction and decision are iterated in successive mergings until the whole image support is represented by a single region.

The tracking of successive merging leads to an optimized BPT image representation spanning a set of mergings from the initial partition (leave nodes) to the root of the BPT.

2.3 Combining Multiple Rules

The main difficulty of using higher level features is how to deal with multiple merging rules. For this purpose we estimate the distribution of the dissimilarity values associated to each rule over the whole image. The idea behind this, is to put in context the significance of the local dissimilarity values computed over pairs of regions.

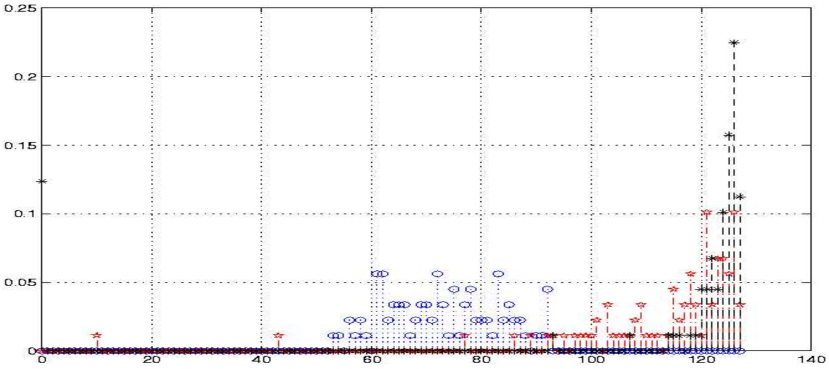

Fig. 3. Pdf estimation for the image shown in Fig. 8. Colour, size and partial inclusion are represented respectively by circles, starts and asterisks.

To estimate the probability density function (pdf) we compute the histogram and divide by the total number of measures. The values of the dissimilarity measures are normalized to the interval $[0, 1]$, and quantified with a sufficient number of bins using a non-linear quantizer. In order to give more relevance to the lower dissimilarity values, we expand this part of the axis and compress the upper part, since lowest values represent rules which tend to be true.

We assume that a **uniform distribution of the dissimilarity values of a rule do not provide relevant information regarding this rule**.

For this purpose we estimate the information of each rule by computing its entropy as, $H_r = -\sum_i p_i log_2(p_i)$. Where p_i is the probability of occurrence of a dissimilarity measure for a given rule. The entropy is the statistical descriptor being computed, and allows combining the individual dissimilarity values in the k-*the* iteration associated to a pair of regions (i, j) in a unique weighted dissimilarity value as,

$$D^k(i, j) = \sum_c w_c^k d_c^k(i, j)$$

where, $w_c^k = \frac{H_c^{max} - H_c}{\sum_{\forall ruler} H_r^{max} - H_r}$.

Note that $\sum_c w_c = 1$ and since $d(i,j) \in [0,1]$ the combined value is also in the interval [0,1]. Entropy is maximal for uniform distributions, thus w_c^k tends to zero decreasing the contribution of this rule in such case.

For example, the image shown in the first row of Fig. 8(a) is over-segmented to provide a set of 49 regions, see 8(b). In this case, there are 89 connected pairs of regions in the first iteration. The rules are assessed providing the dissimilarity measures which are used to estimate the pdf for the aforementioned rules (colour, size, and partial-inclusion). For the initial partition, the distribution of the partial-inclusion dissimilarity values presented in Fig. 3, shows that regions are either totally included or not included.

The distribution of the colour and size values are flatter than the partial-inclusion pdf and the entropy for colour, size and partial inclusion is respectively, 5.0, 4.74 and 3.34. Consequently, their associated weights are $w_{Colour} = 0.25$, $w_{Size} = 0.28$ and $w_{Inc} = 0.46$. As a result, the combined dissimilarity value is weighted so that partial-inclusion is more relevant than size, and size is more relevant than colour. For each pair of connected regions from the current partition a combined dissimilarity measure is computed and the pair of regions with the lowest value is selected for merging.

The syntactic framework allows the combination of simple and complex homogeneity criteria using global statistical information.

2.4 Top-down: Semantic BPT analysis

In the previous motivation section, it has been stated that multiple detections of a single instance is a common problem in image analysis applications based on BPTs. This paper proposes a technique to cope with these cases and choose among all the BPT nodes which are candidate to contain a semantic instance of a class.

Before formulating the selection criterion, it is necessary to introduce the concept of **BPT Semantic Neighborhood**. A BPT Semantic Neighborhood is a subset of connected BPT nodes that represent instances of the same semantic class. Notice that a BPT Semantic Neighborhood is associated to a specific semantic class and that a BPT node could represent more than an instance of different classes. Figure 4 shows and example of two BPT semantic neighborhoods of the same class “A”.

Simple classes We define as **simple** classes those ones whose instances can be completely represented by a single BPT node. They are modelled according to the perceptual caracteristics of the region, for example, as a list of low-level descriptors.

In the simple classes case, the selection criterion is based on the supposition that all instances of the class associated to a BPT Semantic Neighborhood represent in fact the same instance. Among all the candidates, the most similar one to the perceptual model will be chosen and the rest discarded.

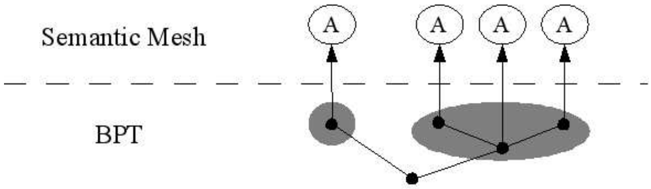

Fig. 4. BPT semantic neighborhood.

Composite classes We consider **composite** classes those which are defined as a combination of semantic instances (SI) of other classes that satisfy certain semantic relations (SR). The model of a composite class includes instances of lower level classes, which, at the same time, are described by other simple and/or composite models. As shown in Fig. 5 a), the semantic decomposition can be iterated until reaching the lowest possible level, corresponding to simple classes. Such top-down expansion can be summarized in a basic graph called Semantic Tree (ST), as shown in Fig. 5 b).

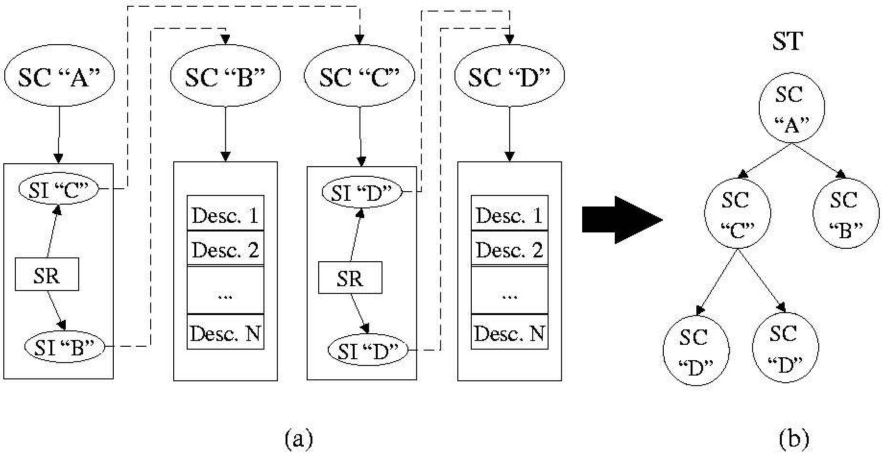

Fig. 5. a) Semantic decomposition, b) Semantic Tree (ST).

When working with composite classes, the detection algorithm will look for combinations of the detected instances that satisfy the relations represented by the model. For each valid combination, a new ST node candidate is created and linked to those ST nodes representing the composing instances. The hierarchical decomposition of classes drives to a recursive detection algorithm [7] according to a bottom-up expansion of the ST.

Towards the final goal of building instances of Semantic Trees, the algorithm prevents creating unnecessary links among ST nodes. Every time a new ST node candidate is added to the Semantic Mesh, it must become the root of the inferior tree structure. That is, the addition of a new node must not close any cycle through the lower levels of the Semantic Mesh. This condition must also be assessed through the BPT semantic neighborhoods. Figure 6 shows a case in which the ST nodes in gray are discarded due to a cycle through a BPT Semantic Neighborhood of class “B”.

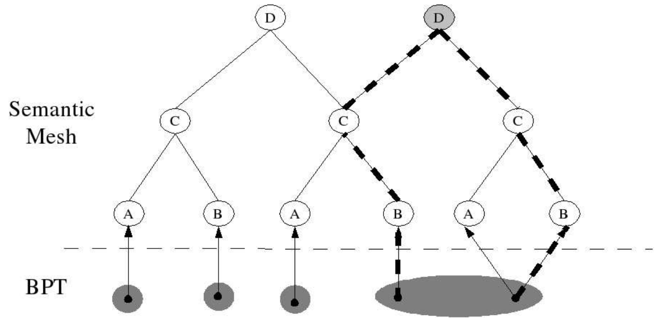

Fig. 6. Cycles discard ST nodes candidates in gray.

The leaves of the STs will always represent simple classes associated to a BPT node, so conflicts similar to the simple class cases may appear. Different BPT nodes in the same neighborhood may be associated to the leaves of different Semantic Trees. Applying the same criterion, this situation represents a conflict and only one instance of a given class can exist in a BPT semantic neighborhood. Our proposal is that when two or more possible ST instances share a BPT semantic neighborhood, we keep the ST instance whose root is in a higher level and discard the remaining ones. This criterion solves the conflict by giving more credibility to the most complex structure as we consider it a good indicator for valid detection. Secondly, when the conflict is among Semantic Trees of the same height, the ST instance most similar to the model of the class is kept. Notice that discarding model is the one represented at the root of the Semantic Tree. Figure 7 shows an example in which the Semantic Tree with probability f_2 is kept and the other two discarded.

3 Results

This section presents two applications of the enhanced BPT framework. Firstly, an example of traffic sign detection illustrates in detail the framework. Secondly, the proposed framework demonstrates its suitability for a general application performing the detection of laptops in a smart room.

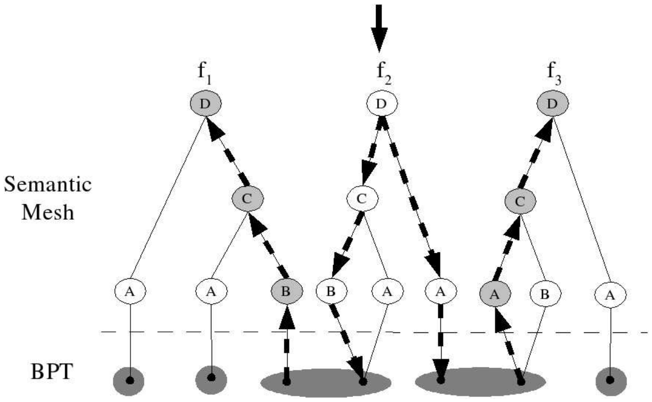

Fig. 7. Consolidated ST ($f2 > f1$ and $f2 > f3$) overlaps with nodes in gray through the BPT semantic neighborhood and forces their deletion.

For both applications the initial partition was automatically created using a colour homogeneity criterion in the YUV space with weights 1, 0.5 and 0.5, respectively and a PSNR of 23 dB for road sign detection and 24dB for laptop detection.

3.1 Application 1: Roadsign detection

The results presented in this application were generated from images from the CLIC database [8] that contains a traffic sign.

The images shown in Fig. 8 present three examples of object extraction using the enhanced BPT framework. As we can see, column a) is the original image, column b) its partition represented by the mean colour of each region and column c) the only detected instance of the object of interest. These examples show the performance of the system for this application.

The detailed analysis of the image shown in the first row is used to exemplify the syntactic and the semantic analysis in the enhanced framework.

Syntactic analysis The aim of this section is to show, by means of a visual example how well the limitations of simple homogeneity criteria, like colour, can be overcome when using the syntactic framework for the BPT creation. For this task, we aim to compare the best object candidate represented as a single node of the BPT that can be obtained using either simple or syntactic homogeneity criteria.

The image on the first row of Fig. 8 has been segmented into 49 colour homogeneous regions. Starting from this initial partition, the BPT is created using only simple homogeneity criteria (colour and size). Figure 9 shows a subtree of this BPT and those nodes representing the best potential object candidates that could be obtained. Although node 85 is a good representation of the inner part

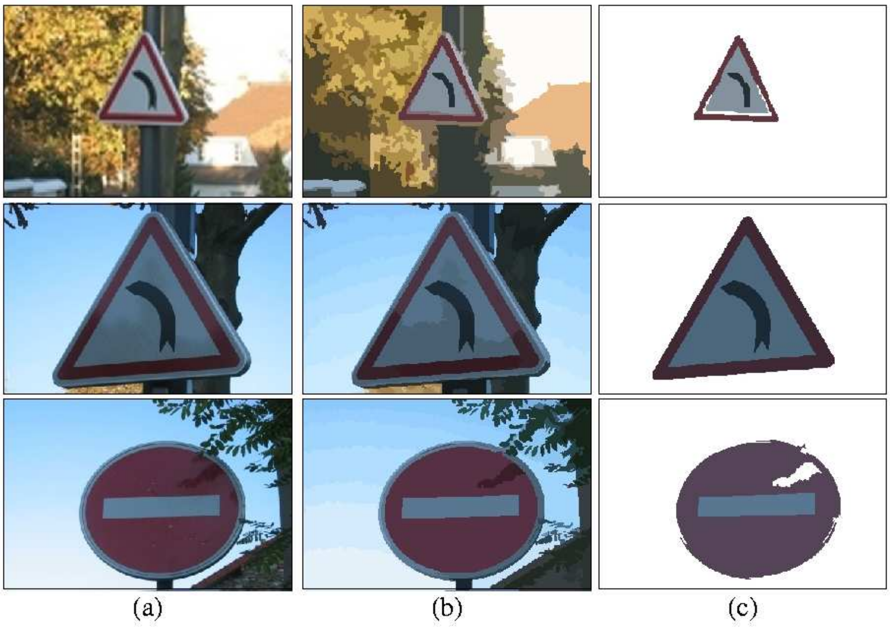

Fig. 8. Application 1: a) Original image, b) Initial partition and c) Extracted traffic sign.

of the sign, the outer part, represented by node 92, has been merged with the background. Therefore, it is not possible to obtain the sign as a single node or as composition of nodes by means of thetop-down analysis using simple homogeneity criteria. Node 93 shows the region resulting from merging the above regions. Neither a descriptor of the whole sign nor a description based on the object parts can detect this object.

In Fig. 10 it is shown the syntactic BPT obtained from the same image partition. This BPT is computed using a combination of simple (colour, size) and complex (syntactic: partial inclusion) criteria. In this case, the sign is represented by a single node which is a better object candidate than the ones found in 9.

Semantic analysis The syntactic analysis created a BPT that, apart from having a node that fully represented the object of interest, it also contained nodes with the parts that composed them. The traffic sign is a type of object that suits its modelization from its parts. The red frame, white background and black silhouette were described separately with colour and shape descriptors, that is, with a perceptual model. Afterwards, references to this classes were used for a semantic modelization of the traffic sign class with a Description Graph [9], as shown in Fig. 11.

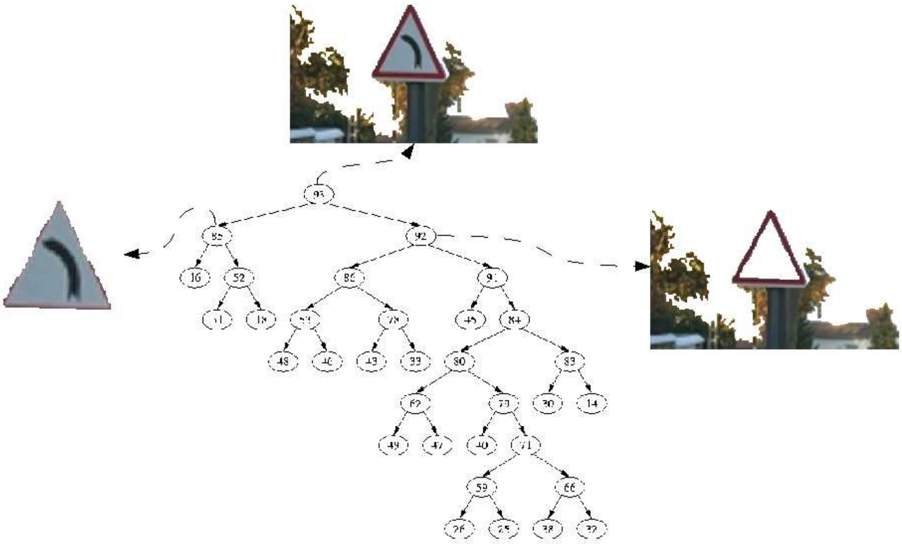

Fig. 9. Application 1: BPT using a simple homogeneity criteria: colour.

Figure 10 a) shows the syntactic BPT created in the previous section. In Fig. 10 b) there are the regions associated to each BPT node. As the perceptual model of the triangular classes *Frame* and *Background* are based on similar shape descriptors, nodes 14, 61, 62 and 63 are a BPT Semantic Neighborhoods for these classes. Nevertheless, the presented technique provided the algorithm with a criterion to resolve the set of candidates that best matched the semantic model. Figure 8 c) depicts the extracted traffic sign composed by its parts.

3.2 Application 2: Laptop detection in a smart room

The flexibility of the algorithm has also been tested in the context of a smart room. The presented technique was applied for the analysis of a QCIF video sequence acquired by a zenital camera and postprocessed with a mask of the table.

The bottom-up analysis leading to the BPT representation is the same as in the previous example and it is performed at each frame of the sequence. The top-down analysis has been adapted by including the desired laptop model. As we can see on Fig. 12, two DGs were used to model the laptops as combinations of screen, chassis, mouse and keyboard. The dual model was chosen to cope with the important variability introduced by the poor image resolution.

Fig. 13 shows three examples of laptop detection in the smart room. Column a) is the original image, b) its associated partition represented by the mean colour of each region and c) is the image with the detected laptops. As we can

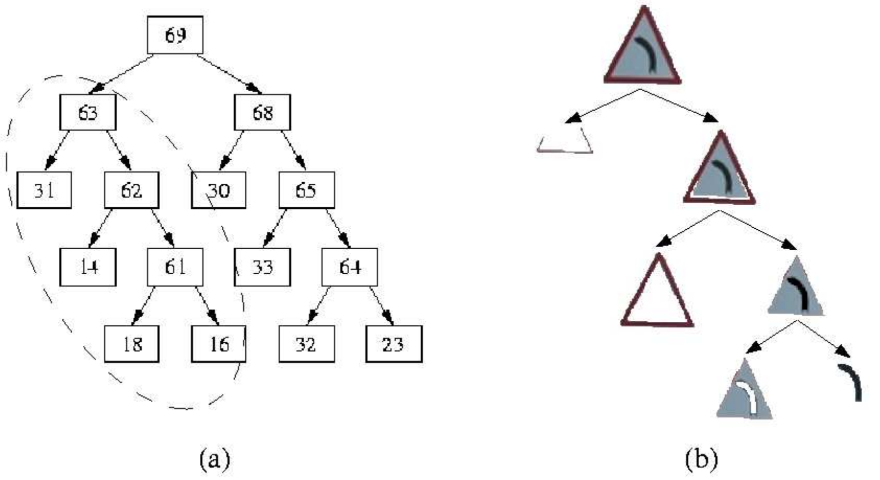

Fig. 10. Application 1: a) BPT using multiple criteria: colour, size and partial inclusion, b) Regions associated to the nodes in ellipse.

see, the first row presents an example where three of the four laptops have been detected. Although the screen has been detected the bottom left laptop is missed because the initial partition has merged the chassis with the body of the person. Including the mouse in the model improves the confidence in the detection of the bottom right laptop and results in a more certain detection. The second row shows the detection of two laptops using two different description graphs. One of them is based on a chassis and a screen, while the second one is based on the keyboard and the screen. In the second case, the chassis is not detected due to an heterogeneous color on the laptop surface. Notice that screens have been correctly segmented despite the narrow visibility angle and a low contrast compared to the chassis. The last row shows a detection of a laptop with a partial occlusion of the keyboard. In this case, the segmentation of the screen was pretty bad, creating a region with a distorted shape. Nevertheless, its proximity to the keyboard is enough to consider it a screen in the context of a laptop.

4 Conclusions

The representation of images from the region-based perspective offered by BPTs provides several advantages for analysis applications. However, how to create this BPTs and how to process it is a key issue for a good performance. This paper has presented two enhancements for the classic BPT.

Firstly, the classic merging of BPT nodes based on colour has been enriched by allowing to combine, using statistical information, multiple criteria, such as

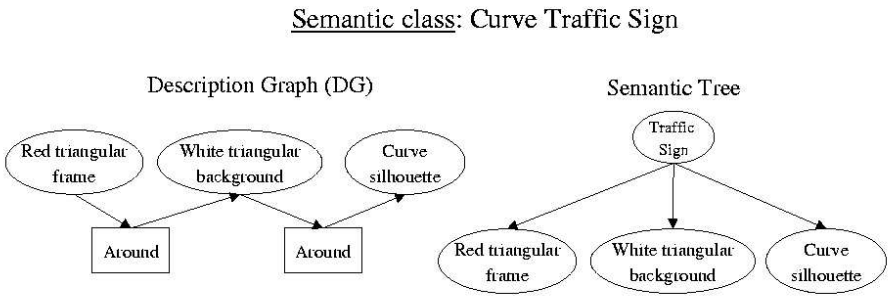

Fig. 11. Application 1: Description Graph and Semantic Tree of the traffic sign.

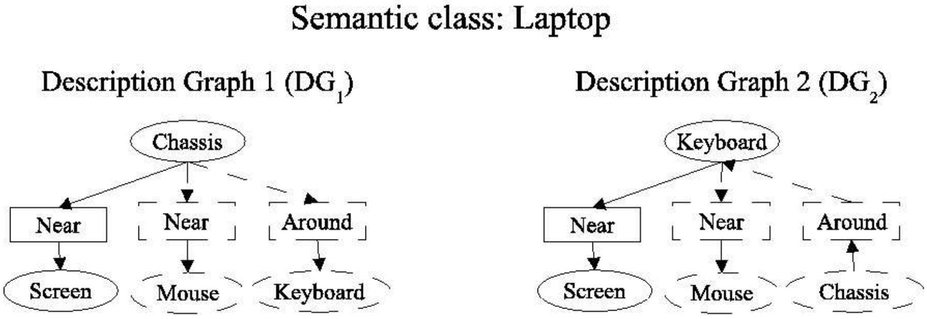

Fig. 12. Application 2: Description Graph and Semantic Tree of the laptop.

the syntactic criteria. A practical example has shown a case in which considering colour, size and partial inclusion of regions drives to the creation of better BPTs.

Secondly, the lack of semantic knowledge of the BPTs usually creates more nodes than semantic entities in the content. Some previous assumptions from the semantic point of view have been presented in this paper based on the definition of BPT Semantic Neighbourhoods. The theoretical approach has also been exemplified on the syntactic BPT, which allowed the detection of a traffic sign from its parts.

The proposed improvements are able to enrich the creation and understanding of BPTs, but can be furtherly expanded. Further research will study new syntactic criteria that will generate new BPTs from the same initial partition. New criteria will may drive to multiple BPTs for the same content, which arises the question of how to combine them from a semantic point of view.

References

1. Salembier, P., Garrido, L.: Binary partition tree as an efficient representation for image processing, segmentation, and information retrieval. IEEE Transactions on

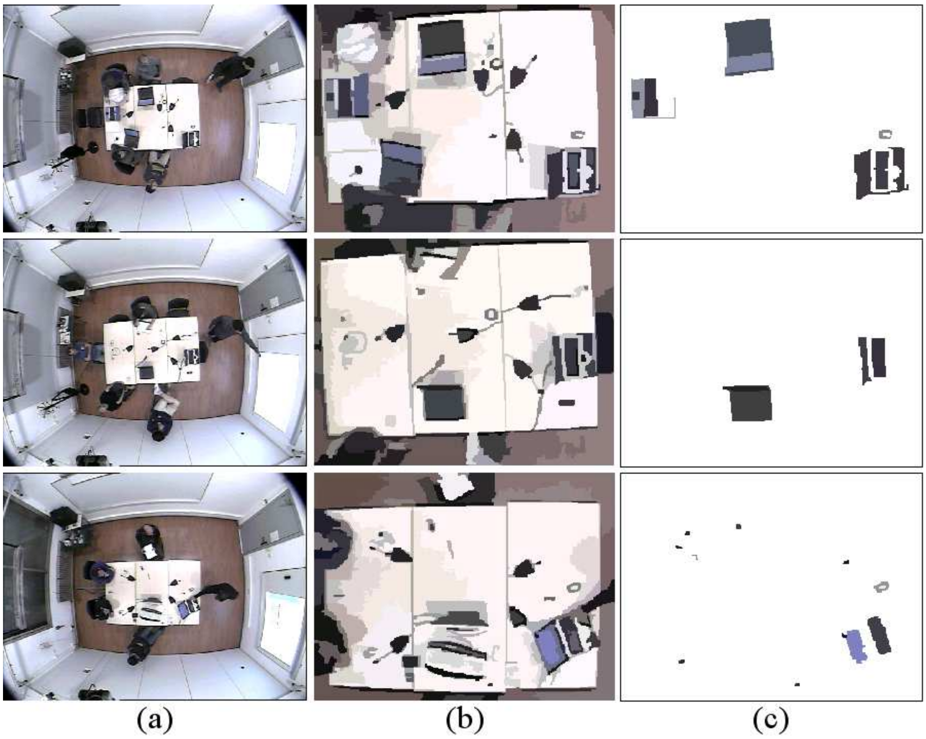

Fig. 13. Applications 2: a) Original image, b) Initial partition and c) Extracted laptop.

Image Processing **9**(4) (2000) 561–576

2. Ferran, C., Casas, J.R.: Object representation using colour, shape and structure criteria in a binary partition tree. In: International Conference on Image Processing (ICIP-2005), Genoa, IEEE (2005) III/1144—1151
3. Koffka, K.: Principles of Gestalt Psychology. (1935)
4. Desolneux, A., Moisan, L., Morel, J.M.: Computational gestalts and perception thresholds. Journal of Physiology (97, Issues 2-3,) (2003) 311–322
5. Fu, K.S.: Digital Pattern Recognition. Springer-Verlag (1976)
6. Garrido, L.: Hierachical Region Based Processing of Images and Video Sequences: Application to Filtering, Segmentation and Information Retrieval. PhD thesis, Signal Theory and Communications Department (UPC), Campus Nord, C/ Jordi Girona 1-3 Modul 5 (2002)
7. Giró, X., Marqués, F.: Detection of semantic objects using description graphs. In: IEEE International Conference on Image Processing, ICIP'05. (2005)
8. P.A Moellic, P.Hede, G.C.: Evaluating content based image retrieval techniques with the one million images clic testbed. In: Proceedings of the International Conference on Pattern recognition and Computer Vision. (2005)
9. Giró, X., Vilaplana, V., Marqués, F., Salembier, P.: 7. In: Multimedia Content and the Semantic Web. Wiley (2005) 203–221