Binary Partition Trees for Object Detection

Veronica Vilaplana, *Member, IEEE*, Ferran Marques, *Member, IEEE*, and Philippe Salembier, *Member, IEEE*

***Abstract*—This paper discusses the use of Binary Partition Trees (BPTs) for object detection. BPTs are hierarchical region-based representations of images. They define a reduced set of regions that covers the image support and that spans various levels of resolution. They are attractive for object detection as they tremendously reduce the search space. In this paper, several issues related to the use of BPT for object detection are studied. Concerning the tree construction, we analyze the compromise between computational complexity reduction and accuracy. This will lead us to define two parts in the BPT: one providing accuracy and one representing the search space for the object detection task. Then we analyze and objectively compare various similarity measures for the tree construction. We conclude that different similarity criteria should be used for the part providing accuracy in the BPT and for the part defining the search space and specific criteria are proposed for each case. Then we discuss the object detection strategy based on BPT. The notion of node extension is proposed and discussed. Finally, several object detection examples illustrating the generality of the approach and its efficiency are reported.**

***Index Terms*—Binary partition tree, hierarchical representation, image region analysis, image representations, image segmentation, object detection.**

I. Introduction

MOST object detection strategies rely on the comparison of the content of an image with an object model at different locations, orientations and resolutions [3]. Typically, object detection algorithms scan the image at numerous positions and scales looking for the possible representations of the object in the scene [2], [27], [33]. This detection process requires both a useful object model and a suitable image representation. Ideally, the object model should characterize in a simple manner all the variability of the object to be detected. In turn, the image representation should compact in the smallest possible number of elements all the information in the scene, while being as generic as possible in order to be able to reuse the representation in different contexts (e.g.,: searching in the same image for different objects). Common image representations are: i) pixel-based representations: the image is understood as a set of independent pixels; ii) block-based representations: the image is seen as a set of rectangular arrays of pixels; iii) region-based representations: the image is represented as a set of homogeneous connected components; and iv) compressed domain representation: the image is seen as a set of coefficients of a particular transform domain, such as the DCT domain.

While pixel- and block-based representations are simple to define, they yield a large number of elements to be analyzed. On the contrary, the definition of region-based and compressed domain representations involves complex image processing operations but largely reduces the analysis space. Given this complexity, there is an interest on reusing the representation in subsequent analysis. Therefore, the definition of an image representation for object detection purposes can be seen as finding a compromise between which analysis steps can be done in a systematic and universal way and which ones actually depend on the specific object to be detected.

Region-based image representations (e.g., [23]) are a good framework for solving this compromise. They provide a simplification of the image in terms of a reduced number of representative elements, which are the regions. In a region-based image representation, objects in the scene are obtained by the union of regions in an initial partition. Since an arbitrary partition may contain about a hundred of regions, the number of different possible unions among these regions can be large. Actually, a region-based representation implies a compromise between accuracy (related to the number of regions in the partition) and processing efficiency (related to the number of unions of regions to be analyzed).

One approach to palliate this problem is to reduce the number of possible region unions by proposing the most likely ones given a specific criterion. This is performed by creating a hierarchy of regions representing the image at different resolution levels. Note that in this work we are not referring to multi resolution as a multiple representation in which, at each resolution, the image is represented by a different number of pixels (see [6] for a general reference on multi resolution analysis). In the region-based framework, the multi resolution notion is related to the number of regions at each level. This region-based multi resolution allows analyzing the image at multiple scales (see [23] for a specific reference on region-based approaches). The idea is to have not only a single partition but a universe of partitions representing the image at various resolutions. In this context, object detection algorithms can be driven to only analyze the image at those positions and scales that are proposed by the regions in the hierarchy. Moreover, regions represent areas of support which allow improving the robustness of the estimation of complex features that can be used in the object detection process.

From this set of regions at various resolution levels, an application dependent algorithm can select the most convenient one(s) for its concrete application. The selected region(s) may

Manuscript received August 30, 2007; revised June 17, 2008. Current version published October 10, 2008. This work was supported in part by the EU project NoE MUSCLE FP6-507752 and in part by the Grant TEC2007-66858/TCM PROVEC and the CENIT-2007-1012 I3MEDIA project of the Spanish Government. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Aria Nosratinia.

The authors are with the Department of Signal Theory and Communications, Technical University of Catalonia (UPC), 08034, Barcelona, Spain (e-mail: veronica.vilaplana@upc.edu; ferran.marques@upc.edu; philippe.salembier@upc.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIP.2008.2002841

directly represent the object to be detected or only provide a good approximation. This approximation could be used as an anchor point for launching a refining process leading to the final detected object.

There exist different approaches to hierarchical region-based image representation, usually related to tree data structures. In these structures, tree nodes are associated with regions in the image whereas tree links represent the inclusion relation. In the quad-tree image representation [20], [26] the image is recursively subdivided into four equal rectangular regions (quadrants) and, therefore, each node has either four children or no children (a leaf). As the subdivision strategy does not depend on the image content, many false contours are introduced in the image representation and it is difficult to use it for general object detection tasks.

The min-tree and max-tree representations [24] adapt to the image content because they describe the image flat zones (largest connected components of the image where the image is constant). The leaves of these trees are the regional minima and maxima of the image, respectively, for the min-tree and the max-tree, whereas the remaining nodes represent flat zones and are ordered by the gray level value differences between regions. Therefore, min-tree and max-tree representations are devoted to dark and bright image components. As a result, the regions associated with their nodes may not conform to real objects in the scene. Note, however, that real objects may not coincide with minima or maxima of the image.

The component tree [18] or tree of shapes [4] merges the min-tree and max-tree into a single representation. It allows representing an image in such a way that maxima and minima can be simultaneously handled. Nevertheless, since the resulting tree is still extremum oriented, nodes commonly do not represent objects in the scene.

On the contrary, the Binary Partition Tree (BPT) [22] reflects the similarity between neighboring regions. It proposes a hierarchy of regions created by a merging algorithm that can make use of any similarity measure. Starting from a given partition (that can be as fine as assuming that each pixel or each flat zone is a region), the region merging algorithm proceeds iteratively by 1) computing a similarity measure (*merging criterion*) for all pair of neighbor regions, 2) selecting the most similar pair of regions and merging them into a new region, and 3) updating the neighborhood and the similarity measures. The algorithm iterates steps 2) and 3) until all regions are merged into a single region.

The BPT stores the whole merging sequence from an initial partition to the one-single region representation. The leaves in the tree are the regions in the initial partition. A merging is represented by creating a parent node (the new region resulting from the merging) and linking it to its two children nodes (the pair of regions that are merged). An example of BPT is shown in Fig. 1. As with the other tree representations, the nodes of the tree represent regions and the links the inclusion relationship.

The BPT represents a set of regions at different scales of resolution and its nodes provide good estimates of the objects in the scene. As previously said, classical object detection algorithms scan the image at numerous positions and scales looking for the possible representations of the object in the scene. Using the BPT representation, the image has to be analyzed only at the positions and scales that are proposed by the BPT nodes. Therefore, the BPT can be considered as a means of reducing the search space in object detection tasks.

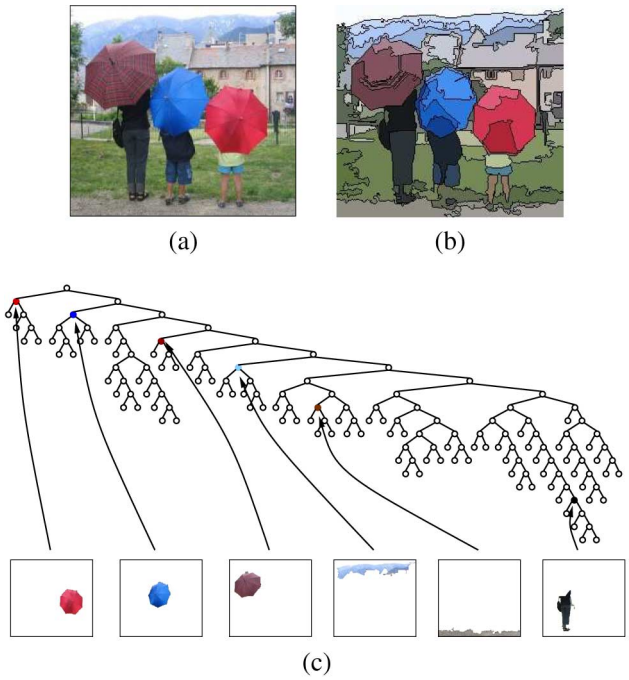

Fig. 1. Example of BPT and illustration of its ability to represent objects in the scene: (a) Original image, (b) Initial partition where each region has been filled with its mean color, (c) Binary Partition Tree and examples of nodes representing objects in the scene.

In this paper, we study the use of BPT for object detection. We propose strategies to provide accuracy (definition of the initial partition) and efficiency (selection of the nodes to be analyzed) to the representation. To this goal, we will highlight two set of nodes in the BPT, a set providing accuracy and another set defining the search space. We also study various similarity criteria that can be used for the BPT construction and objectively assess their performances. Finally, we propose a strategy relying on the notion of *extended node* to perform the detection of objects and illustrate the usefulness of this approach for the specific application of face detection.

This paper is organized as follows. The following section discusses the BPT representation and the associated accuracy and efficiency compromise. Section III presents the methodology that has been used to perform the experiments. Section IV focuses on the BPT construction and, in particular, on the possible similarity criteria that can be used. The generic strategy used for object detection is proposed and demonstrated in Section V. Finally, conclusions are reported in Section VI.

II. Image Representation Based on Binary Partition Tree

The image representation and its main features in the context of object detection are illustrated in Fig. 2. The BPT is constructed from its leaves by successive merging steps. The leaves of the tree form a partition of the space that is commonly named

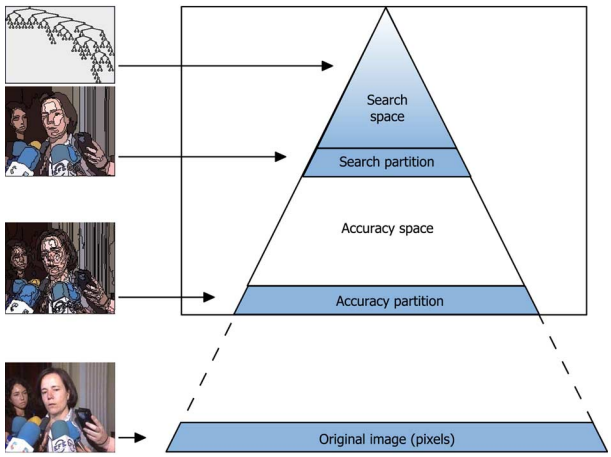

Fig. 2. Image representation based on the BPT.

the initial partition. As mentioned in the introduction, the initial partition can be made of individual pixels or of flat zones. This is, however, neither very useful nor practical in most object detection applications as the size of the resulting BPT is quite large. In most applications, the use of a very accurate partition with a fairly high number of regions is more appropriate. This initial partition can be created by any segmentation algorithm. For the experiments reported in this paper, the initial partition is created by the same merging algorithm as the one used to create the BPT (see Section IV). If this merging algorithm is used to define a partition, a *stopping criterion* is assessed at each merging step. Since in the context of object detection this partition is used to ensure an accurate representation of the objects in the scene, this partition is called in the sequel the *accuracy partition* (see Fig. 2).

The BPT represents the sequence of merging of neighboring regions assuming that the most similar pair of regions is merged first. This image representation is a hierarchical region-based representation of the image. At lower scales, that is, at scales close to the initial partition, one will find a very large number of small regions. This set of nodes is important to provide information about the image details and eventually accuracy to the image representation. As regions are merged, higher scales are created. They represent regions of the image that are progressively larger and possibly more meaningful. Note that, as the region size increases, it becomes possible to robustly measure some of the region features and to analyze them.

In the context of object detection, it is useless to analyze very small regions because they cannot represent meaningful objects and, furthermore, measurements will not be reliable. Moreover, the number of very small regions may be fairly high and if the application involves severe restrictions on the computational complexity of the object detection scheme, the search space should be as small as possible. As a result, we distinguish two zones in the BPT: the *accuracy space* providing preciseness to the description (lower scales) and the *search space* for the object detection task (higher scales). A way to define these two zones is to specify a point of the merging sequence starting from which the regions that are created are considered as belonging to the search space. A specific point of the merging sequence is obtained by assessing a stopping criterion. The partition that is obtained at this point of the merging process is called the *search partition* (see Fig. 2).

Within the framework of the BPT representation, the work in [15] adopts a different approach to tackle the trade-off between accurate description and usefulness for object detection. In it, a BPT is created having as initial partition the result of a watershed (that is, a partition containing a large number of regions, typically around 10.000). Then, the evolution of the merging criterion is analyzed at every branch of the BPT. Based on the so-called reluctancy of the merging, subbranches are removed from the representation; that is, the BPT is locally simplified. Although this representation has been proposed as basis for semantic object extraction [16], from our perspective, the simplification approaches proposed in [15] do not fulfill the requirements to define either an accuracy or a search partition (reported values for the so-called conservative algorithm lead to partitions of around 300 regions, whereas those for the so-called bold algorithm lead to partitions of less than ten regions).

Section IV discusses the creation of the image representation exemplified in Fig. 2. It focuses in particular on how to define the accuracy and the search partitions as well as on the possible similarity and stopping criteria for the merging steps.

III. Methodology for the Experiments

To assess the quality of the representation, we will present in Section IV various criteria to build the tree and we will analyze them in terms of partition-based metrics. Throughout the paper, if the proposed experiment allows it, we use three different databases for assessment purposes. The objects in these databases have been manually segmented and the resulting object partitions are used as ground truth in the experiments.

- To analyze generic features, a set of 100 images from the ©Corel image database is used. The set contains ten images of ten different complexity classes which are grouped in the following way: *tigers, horses, eagles, mountains, fields, cars, jets, beaches, butterflies* and *roses*. The objects in this Corel subset (160 in total) have been manually segmented in the context of the SCHEMA project (http://www.iti.gr/SCHEMA/).

To analyze specific applications, two additional databases are used.

- A set of 100 images from the MPEG-7 database [19] has been selected. These images contain human faces in scenarios of different complexity that have been manually segmented leading to a total of 116 objects.
- A set of 45 images from a traffic sign database [14] has been selected. The set contains 15 images of three different types of sign shapes (square, triangle and circle). They have been manually segmented leading to a total of 45 objects.

There exist several proposals for comparing image partitions [8], [31]. Most of these techniques compare two simple partitions (that is, partitions with only a few regions) where each region represents an object in the scene. This type of partition is commonly named object partition. In our case, in addition to this type of partition comparison, we want to asses how well the regions in a dense partition (that is, a partition with a large number

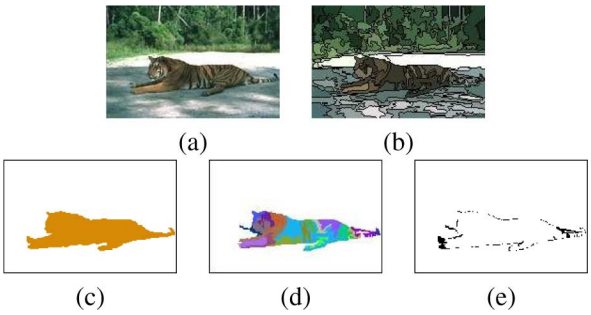

Fig. 3. Illustration of asymmetric distance: (a) Original image. (b) Example of search partition (regions filled with their mean value). (c) Object partition. (d) Regions from the search partition partially contained in the object region. (e) Pixels requiring a label change.

of regions) accommodate to the shape of the objects represented in a ground truth object partition.

We have selected the approach proposed in [7] since it provides a single framework for both cases. Initially, a symmetric distance is proposed in [7], $d_{\mathrm{sym}}(P,Q)$, which is defined in terms of the number of pixels whose labels should be changed between regions in P to achieve a perfect matching with Q (P and Q become identical). In our work, the symmetric distance will be used for comparing two object partitions (P,Q) (for instance, a ground truth object partition and the object representation obtained by selecting a node in the BPT, as in Section V).

The definition is extended in [7] to an asymmetric distance, $d_{\mathrm{asy}}(P,Q)$, in a way such that the distance between a partition P and any partition Q, finer than P, is zero. A partition Q is said to be finer than a partition P if and only if the intersection of P and Q is equal to Q. Therefore, the asymmetric partition distance is defined as the minimum number of pixels whose labels should be changed so that partition Q becomes finer than partition P. In our work, the asymmetric distance will be used for assessing how well an object partition (P) can be matched by the result of merging regions from a denser partition (Q) (for instance, a ground truth object partition and a search partition, as in Section IV).

In [7], the number of label changes for both distances is normalized by the image size to obtain the final distance values. In our work, given that we are always using a ground truth object partition, we can normalize the distance values using the object size. An example of how the asymmetric distance is computed is presented in Fig. 3. Partitions shown in Fig. 3(b) and (c) are compared in terms of the number of pixels whose labels should be changed so that partition (b) is finer than (c). Fig. 3(e) shows the pixels that require a label change.

IV. Creation of the BPT

A. Merging Criteria and Region Model

The creation of a BPT relies on two major notions: the merging criterion and the region model [22]. The merging criterion defines the similarity of neighboring regions and, therefore, the order in which regions are going to be merged. The region model specifies how regions are represented and how to compute the model of the union of two regions.

In this work, the region model M_R is assumed to be constant within the region R, and is the vector formed by the average values of all pixels $p \in R$, in the YCbCr color space

$$M_R = \frac{1}{N_R} \sum_{p \in R} I(p) \tag{1}$$

where N_R is the number of pixels of region R.

The similarity measure or merging criterion is computed for each pair of neighboring regions, R_1 and R_2, according to a selected homogeneity criterion. The basic criterion used in most still image segmentation approaches is color homogeneity. Some of the measures are size independent, like the mean squared error (MSE) between the merged region and its model [12]

$$O_{\mathrm{MSE}}(R_1,R_2) = \sum_{p \in R_1 \cup R_2} (I(p) - M_{R_1 \cup R_2})^2 / (N_1 + N_2) \tag{2}$$

or the Euclidean distance (ED) between the region models [34].

In general, size independent color-based measures tend to produce partitions with few large regions and a large number of extremely small regions. Trying to avoid this problem, other measures have been proposed taking into account the region sizes, as the squared error (SE) [12]

$$O_{\mathrm{SE}}(R_1,R_2) = \sum_{p \in R_1 \cup R_2} (I(p) - M_{R_1 \cup R_2})^2 \tag{3}$$

or the weighted Euclidean distance (WED) [1]. When using these size dependent criteria, the cost of merging for small regions decreases, forcing small regions to merge together first and encouraging the creation of large regions.

As a compromise between the two types of measures, two other measures are analyzed here. They compare the models of the original regions with the model of the region obtained after the merging [12]. The weighted squared distance between region models (WSDM) is defined as

$$O_{\mathrm{WSDM}}(R_1,R_2) = N_{R_1} \|M_{R_1} - M_{R_1 \cup R_2}\|_2^2 + N_{R_2} \|M_{R_2} - M_{R_1 \cup R_2}\|_2^2 \tag{4}$$

and the weighed Euclidean distance between region models (WEDM) as

$$O_{\mathrm{WEDM}}(R_1,R_2) = N_{R_1} \|M_{R_1} - M_{R_1 \cup R_2}\|_2 + N_{R_2} \|M_{R_2} - M_{R_1 \cup R_2}\|_2. \tag{5}$$

Finally, and based on experimental evidences, we observed that while luminance information is crucial to define visually relevant contours, chrominance information is paramount in the definition of objects. Since our final goal is to have regions in the search area being as close as possible to objects in the scene, we introduce another distance modifying the WEDM by increasing the relevance of chrominance information. We also introduce a second change based on the use of contour information. Since most 'real' objects are regular and compact (that is, tend to have simple contours), the analysis of shape complexity can provide additional information for the mergings. Therefore, we include

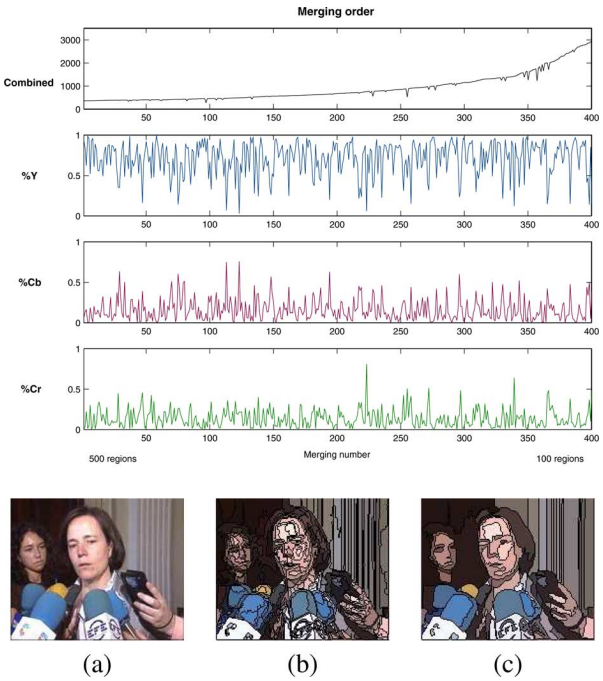

Fig. 4. Relevance of the image components and evolution of region sizes during the merging process using the WEDM. Original image (a) and partitions with (b) 500 and (c) 100 regions.

in the similarity measure a term related to the contour complexity of neighboring regions.

As a consequence, the proposed merging criterion (which implies a normalized weighted Euclidean distance between models with contour complexity and, for simplicity, is referred to as NWMC in the sequel) has two terms. One term is based on color similarity. The color difference in each component is normalized by the dynamic range of the component in the image. This way, it adapts to the chrominance variability of the image. For each image component, we compute the weighted Euclidean distance between models normalized by the component dynamic range

$$O_{\text{color}}(R_1,R_2) = N_{R_1}\|w(M_{R_1}-M_{R_1\cup R_2})\|_2 + N_{R_2}\|w(M_{R_2}-M_{R_1\cup R_2})\|_2 \quad (6)$$

where w is a vector being w_i the inverse of the dynamic range of the image component $i \in \{Y, Cb, Cr\}$ (that is, the difference between the Max value and the Min value of the image component).

The second term is related to the contour complexity of the merged regions. After analyzing several approaches, the adopted measure computes the increase in perimeter of the new region with respect to the largest of the two merged regions: $\Delta P(R_1,R_2) = \min(P_1,P_2) - 2P_{12}$, where P_1 and P_2 are the R_1 and R_2 perimeters, respectively, and P_{12} is the common perimeter between the regions. The term that measures contour complexity is

$$O_{\text{cont}}(R_1,R_2) = \max(0, \Delta P(R_1,R_2)) \quad (7)$$

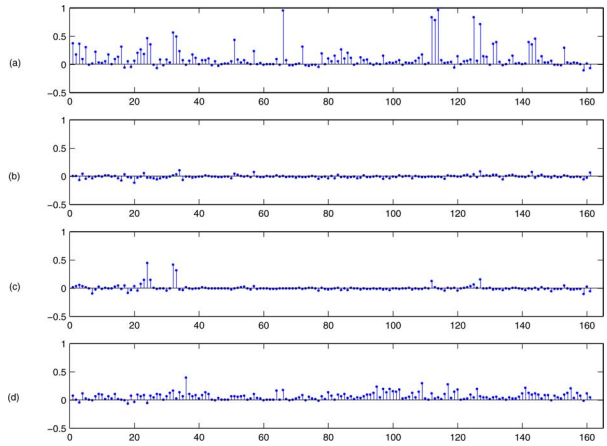

Fig. 5. Difference between the asymmetric distance values computed for different merging criteria on the COREL subset. (a) MSE-WEDM. (b) SE-WEDM. (c) WSDM-WEDM. (d) NWMC-WEDM.

which sets to 0 negative increments that occur when a region is partially or totally included in the other. Color and contour similarity measures are linearly combined to form the NWMC

$$O_{\text{NWMC}}(R_1,R_2) = \alpha O_{\text{color}}(R_1,R_2) + (1-\alpha)O_{\text{cont}}(R_1,R_2). \quad (8)$$

In the sequel, these distances will be investigated in the context of BPT construction for object detection. Note that in the NWMC, there is a parameter α that has to be set. The sensitivity of this parameter will be analyzed at the end of this section. As it will be shown (see Fig. 14), the sensitivity of the segmentation process is very low with respect to variations of the alpha parameter around $\alpha = 0.5$ and, therefore, this value is used throughout the sequel.

B. Definition of the Merging Criterion for the Accuracy Area

In Sections IV-B and IV-C, we analyze the merging and stopping criteria, respectively, to create the accuracy area and the search partition. In this analysis, the starting point for the merging algorithm is the pixel level. Almost all merging criteria that are proposed could be applied to define a useful accuracy partition using a simple stopping criterion (such as reaching a fair number of regions; typically, 500 regions). For simplicity, we will discuss the creation of the accuracy area and search partition (Experiment 1–3) and, once the merging criterion will be selected, we will assess the quality of the resulting accuracy partition (Experiment 4).

Experiment 1: To quantify the performance of the similarity measures, different merging criteria are compared on the 100 images subset of the Corel database (see Table I). In this experiment, to decouple the effects of the merging and stopping criteria (that is, of the creation of the accuracy area and the search partition), a simple stopping criterion is used: merge up to 50 regions. The comparison is performed in terms of mean values of the final PSNR (between the image created by filling each region with its model and the original image) and of the variance of the region sizes.

In terms of PSNR, the WEDM criterion improves all criteria except the WSDM which is only slightly better. However, these

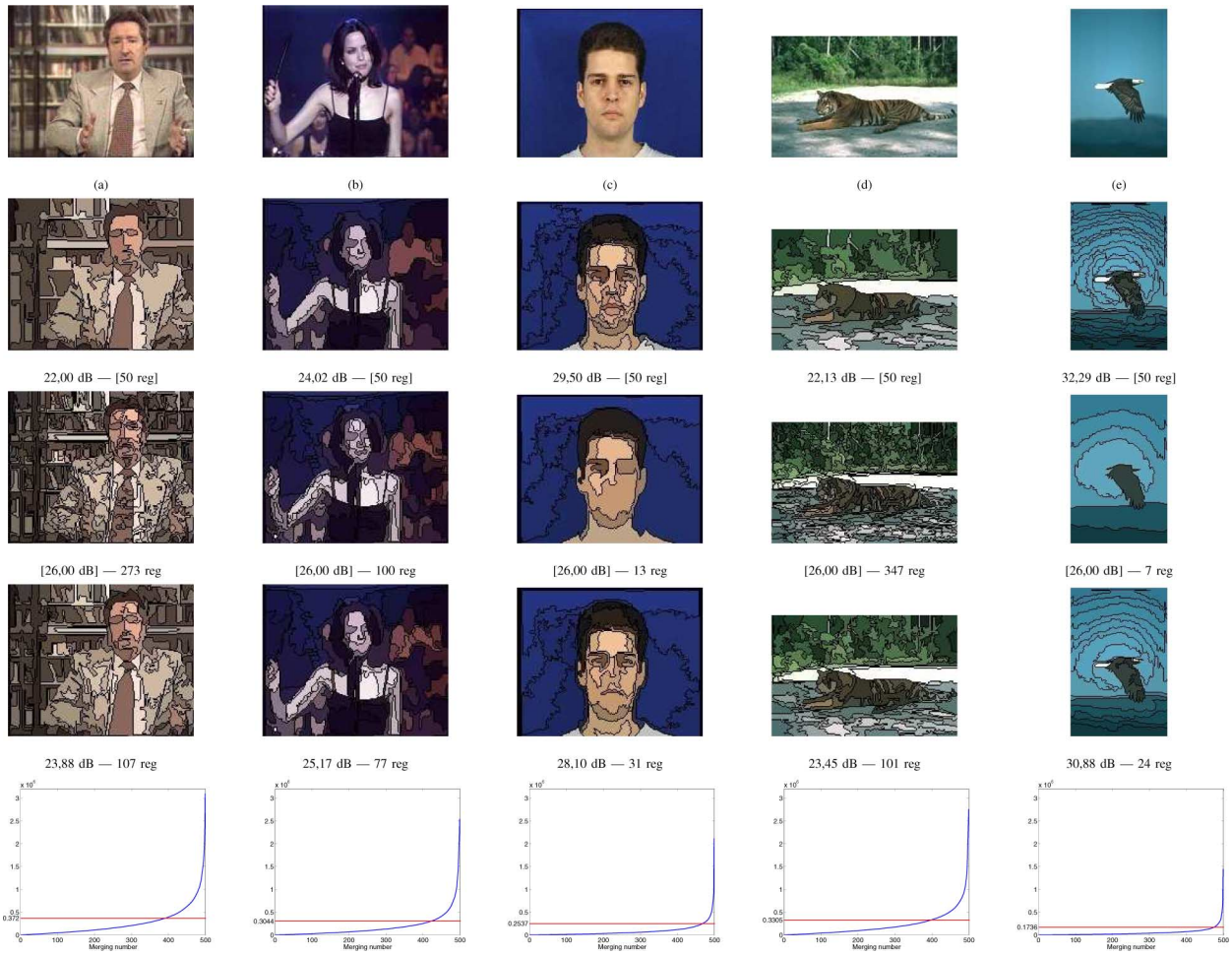

Fig. 6. Comparison of the different stopping criteria. First row: original images. Second row: SC1: $N_{\mathrm{reg}} = 50$. Third row: SC2: $\mathrm{PSNR} = 26$ dB. Fourth row: SC3: $T_{\mathrm{ACM}} = 0.12$. Fifth row: AMC(m).

TABLE I
MERGING CRITERIA COMPARISON ON THE COREL SUBSET. PSNR VALUES ARE GIVEN IN DECIBELS. σ^2_{reg} VALUES ARE DIVIDED BY 10^5

	MSE	SE	WSDM	WEDM	NWMC
PSNR	22,99	24,72	25,33	24,87	20,16
σ^2_{reg} **on region size**	16,90	2,86	4,74	1,51	2,18

results for the WEDM are obtained while largely outperforming the other criteria in terms of variance of the region sizes. As previously commented, this is a relevant feature since it ensures that regions obtained with the WEDM will be adequate for a subsequent robust feature estimation: large enough and homogenous in size while presenting similar PSNR values than previous measures and leading to visually good representation. This behavior is illustrated in the second row of Fig. 6 with five images of different complexities. Regarding the NWMC criterion, note that it yields the worst results in terms of PSNR. This is due to the fact that this criterion tends, as commented previously, to create regions with smooth contours which are not common in the lower scales of the image representation.

Fig. 4 illustrates the behavior of the WEDM as well as the importance of each image component in (5). In this example, the plots show the evolution of the similarity values for the merging process starting from an accuracy partition containing 500 regions [see Fig. 4(b)] up to achieving 100 regions [see Fig. 4(c)].

The plot on the top shows the value of the global merging order; that is, combining the three image components as described by (5). The next three plots show the percentage of each component in the global measure given by (5). Note that most of the mergings are performed based on luminance similarity (the luminance percentage is high) and the contribution of the chrominance components remains quite stable through the merging sequence.

Experiment 2: In this experiment, the quality of the partitions obtained under the conditions of Experiment 1 is assessed in terms of their accuracy representing objects. We use the 160 objects manually segmented from the Corel database subset. The asymmetric distances between the partitions obtained with each of the five merging criteria (using as stopping criterion $N_{\mathrm{reg}} = 50$ regions) and the object partition (ground truth) are computed for each object. Fig. 5 shows the difference between the asymmetric distance values obtained using the four previous merging criteria and the WEDM merging criterion for each one of the 160 objects of the Corel database subset. Statistics of each merging criterion are presented in Table II. Note that, in this case, the global behaviors of the SE and WEDM criteria are

TABLE II
MERGING CRITERIA COMPARISON ON THE COREL SUBSET. ASYMMETRIC DISTANCE MEAN AND VARIANCE VALUES ARE MULTIPLIED BY 10^2

Asymmetric distance	MSE	SE	WSDM	WEDM	NWMC
Mean	22.57	10.52	11.46	10.53	17.73
σ^2	5.549	1.012	1.880	1.038	1.610

very similar, outperforming those of MSE, WSDM and NWMC criteria.

Summarizing the results of the previous experiments, it can be observed that the WEDM merging criterion largely outperforms the MSE, WSDM and NWMC criteria, while improving the SE one. Therefore, we propose the WEDM as merging criterion for the definition of the accuracy area of the BPT.

C. Definition of the Search Partition: Stopping Criterion

As explained in Section II, we distinguish two zones in the BPT: the accuracy space providing preciseness to the description and the search space for the object detection task. These zones are specified by a stopping criterion that defines a point of the merging sequence or equivalently a search partition that is obtained at this point of the merging process.

Typical stopping criteria deal with reaching an *a priori* value of a parameter such as the final number of regions in the partition or the PSNR between the image created by filling each region with its model and the original image. However, as we are defining the search space above the search partition, the objective is to create regions corresponding to parts of the objects in the scene whilst avoiding the creation of regions spanning more than one object. Thus, the stopping criterion has to take into account the complexity of the scene.

We propose a procedure to estimate this complexity based on the *accumulated merging cost*; that is, on the measure adopted from the set of similarity measures defined in Section IV-A. Let $O(k)$ be the cost of the merging at iteration k; that is, the similarity measure between the regions merged at iteration k. The accumulated merging cost (AMC) is defined as

$$\mathrm{AMC}(m) = \sum_{k=1}^{m} O(k). \tag{9}$$

The stopping criterion is defined as a fraction $T_{\mathrm{AMC}} \in [0,1]$ of the total AMC (which is $\mathrm{AMC}(N-1)$, where N is the number of regions in the initial partition). This criterion stops the merging process at iteration $\bar{m}$, where

$$\bar{m} = \min\{m/\mathrm{AMC}(m) > \mathrm{AMC}(N-1)T_{\mathrm{AMC}}\}. \tag{10}$$

Note that if the similarity measure used in the merging process is the weighted square difference of region models (WSDM) [12] and the initial partition defines each pixel as a different region, then the accumulated cost equals the squared error and the stopping criteria becomes a threshold relative to the maximum PSNR.

A stopping criterion based on the analysis of the accumulated cost was also proposed in [1]. However, this approach is not useful in our case since the method in [1] aims at achieving final partitions with a very reduced number of regions. In turn, as commented in Section II, the work in [15] proposed to analyze the evolution of the similarity measure in the context of the BPT representation. As previously, the final number of regions reported in [15] makes this technique not useful for our purposes (the conservative policy leading to around 300 regions, whereas the nonconservative one leading to less than ten regions).

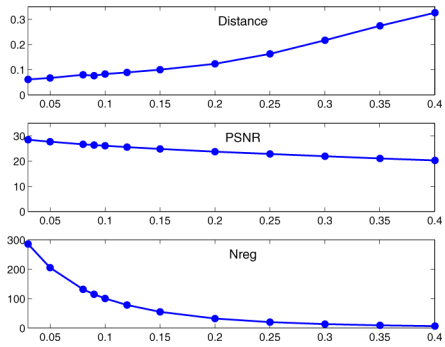

Fig. 7. Analysis of the T_{AMC} impact in terms of mean values of the asymmetric distance, PSNR and number of regions of the partitions obtained with the Corel database subset.

Fig. 6 compares, for a set of images with different complexity, the results obtained by the most common stopping criteria: final number of regions N_{reg}, final PSNR and the proposed AMC criterion. In all cases, the merging criterion is the WEDM measure. The second row presents the partitions obtained using as stopping criterion a fixed number of regions, $N_{\mathrm{reg}} = 50$. In the third row, the final PSNR is fixed to 26 dB whereas the fourth row shows the results for the new criterion, with $T_{\mathrm{ACM}} = 0.12$. As it can be seen, the proposed criterion adapts to the image complexity; that is, it avoids oversegmentation as well as undersegmentation effects, obtaining partitions in which the main objects in the scene are correctly represented.

The last row of Fig. 6 shows the accumulated costs plotted for each iteration of the merging process, starting with an accuracy partition composed of 500 regions. Note that images with different complexity lead to ACM(m) plots of different behavior. Plots also show the thresholds obtained for $T_{\mathrm{AMC}} = 0.12$. This value has been selected after analyzing the robustness of the system using the Corel database subset (in terms of final PSNR, average number of regions and average distance to object partitions) with respect to variations of T_{ACM}, as presented in Fig. 7. As it can be seen, the selected value represents a good compromise among the assessed features since it leads to a low asymmetric distance and a high enough PSNR values while yielding a reasonably small number of regions (less than 100).

Experiment 3: The proposed stopping criterion is quantitatively assessed, as in Section IV-B, by using the Corel subset and the asymmetric distance (see Section III). In this case, the similarity measure is the WEDM and we compare as stopping criteria a fixed N_{reg}, a fixed PSNR and the proposed AMC. The used N_{reg} and PSNR values are the mean values obtained by the AMC criteria over the Corel database subset ($N_{\mathrm{reg}} = 77$ and $\mathrm{PSNR} = 25.54$ dB).

Fig. 8 shows the difference between the asymmetric distance values obtained using the two previous stopping criteria with respect to the AMC stopping criterion for each one of the Corel

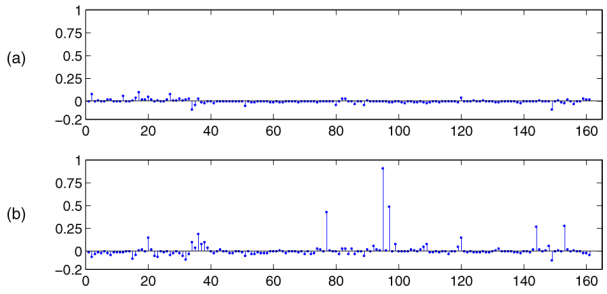

Fig. 8. Difference between the asymmetric distance values. (a) $N_{\text{reg}} - \text{AMC}$. (b) PSNR-AMC.

TABLE III
STOPPING CRITERIA COMPARISON ON THE COREL SUBSET. ASYMMETRIC DISTANCE MEAN AND VARIANCE VALUES ARE MULTIPLIED BY 10^2

	Nreg	PSNR	AMC
Mean	8.89	10.37	8.81
σ^2	0.66	1.70	0.57

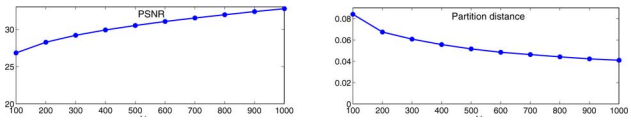

Fig. 9. Analysis of the impact of the number of regions in the accuracy partition.

160 objects. Notice that, for complex images, the WEDM criterion outperforms the N_{reg} criterion. This is the case of the classes *tigers* and *horses* where the amount of regions is too low producing undersegmented results that lead to higher asymmetric distances [see examples in Fig. 6(a) and (d)]. In turn, for simple images, the WEDM criterion outperforms the PSNR criterion. This is the case of the classes *eagles* and *jets* where the selected PSNR is too low producing undersegmented results that lead to higher asymmetric distances [see examples in Fig. 6(c) and (e)].

The statistics of each stopping criterion are presented in Table III. It can be concluded that the AMC stopping criterion (computed over the WEDM merging criterion) outperforms the PSNR criterion, while slightly improving the results obtained with the N_{reg} criterion.

As it has been shown, the behavior of the AMC criterion can be, in several cases, approximated by the N_{reg} criterion. This is the reason why we have proposed to define the accuracy partition by a fixed (and rather large) number of regions. In turn, the ACM criterion is computed starting from this accuracy partition. The accuracy partition represents the highest resolution in the hierarchy and, therefore, should ensure a sufficient definition in the scene representation. We analyze this point in the next experiment.

Experiment 4: The accuracy partition is assessed by segmenting the Corel database subset, using the WEDM merging criterion and the N_{reg} stopping criterion. Fig. 9 presents, for different values of the final number of regions, the mean values of the PSNR corresponding to the accuracy partitions and of their asymmetric distance with respect to the Corel database subset. As it can be seen, the evolution of both parameters is smooth and steady. For the selected number of regions (500), the obtained values are 31 dB and 0.057, respectively, which, as it is shown

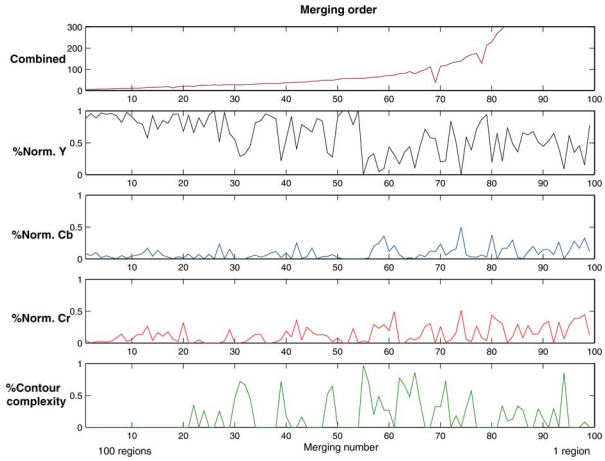

Fig. 10. Example of priority values used at each step during the merging process to build a BPT from a search partition with 100 regions.

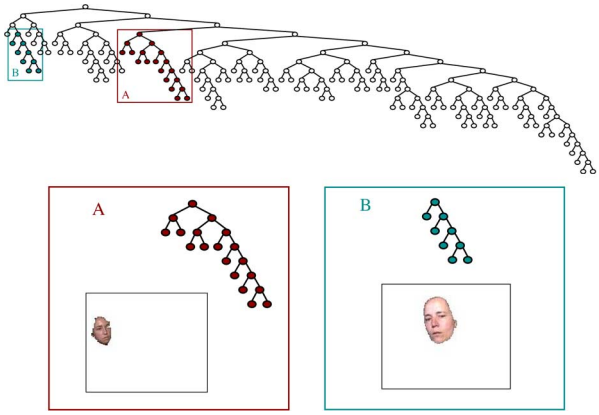

Fig. 11. Search space of the BPT created with NWMC.

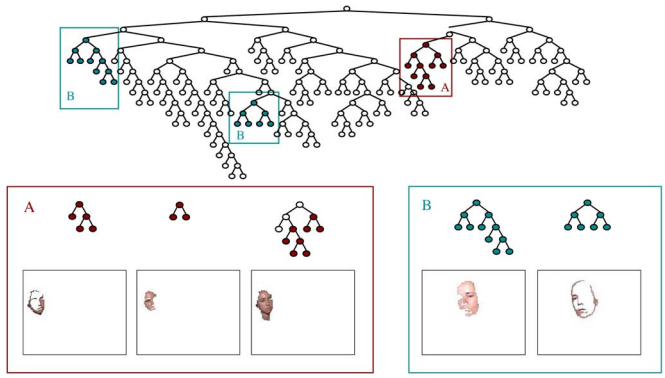

Fig. 12. Search space of the BPT created with WEDM.

through the various experiments in this paper, represent a good compromise.

D. Definition of the Search Space of the BPT

Here, we analyze the construction of the hierarchical structure from the regions defined by the search partition. The discussion is centered on the similarity measure. Ideally, nodes in the tree defining the search space should be objects or parts of objects with a semantic meaning. Therefore, the similarity measure should be related to a notion of object. Several approaches to segmentation try to create 'meaningful' partitions incorporating geometric features into the segmentation

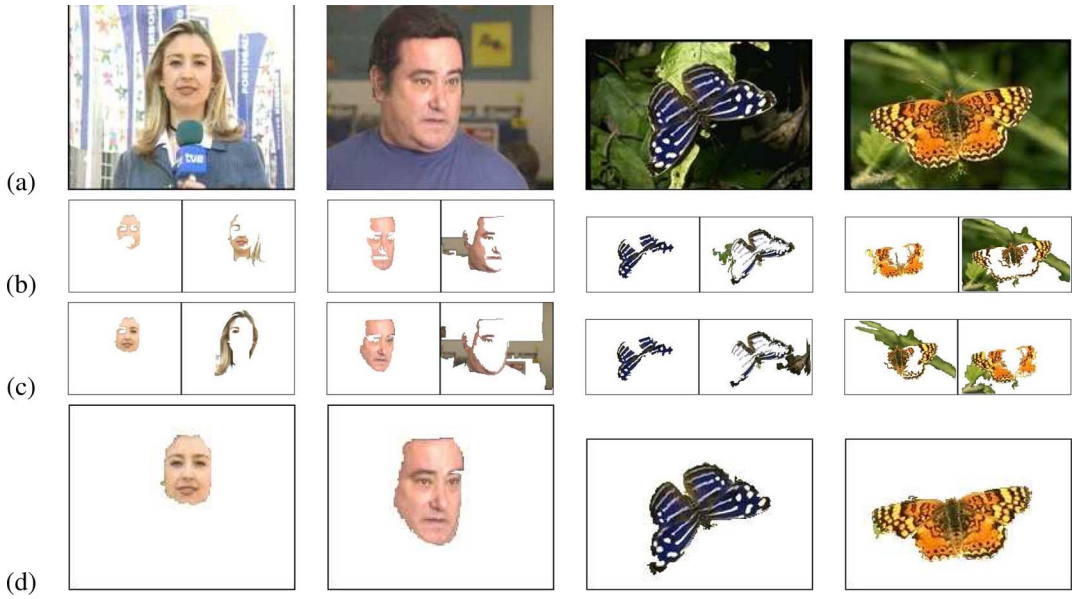

Fig. 13. Face nodes obtained with a criterion based on equally weighted components: row (a) Original images, row (b) WEDM, row (c) a criterion based on normalized components, and row (d) a combination of color and contour complexity: NWMC.

process, like measures of proximity, compactness, inclusion or symmetry (see [1] and [11]). However, the integration of this information is difficult to analyze and evaluate, since there is a strong overlap between various geometrical features (e.g., adjacency, contour complexity and quasi-inclusion). Moreover, in some cases the homogeneity assumptions implied by these features are too strong to lead to generic segmentations; that is, the resulting partitions are already biased to a specific type of object. This analysis led us to the proposal of the NWMC criterion defined in Section IV-A

Fig. 10 shows the priority values for the sequence of mergins performed to create the upper part of BPT for the image and the search partition presented in Fig. 4(a) and (c), respectively, using the NWMC criterion. The plots represent the combined measure, the percentage of the normalized color measures for each component and the percentage of the contour complexity term. The use of the term related to contour complexity favors the merging of regions with partial or total inclusions, unless they are very different in color. The first merging steps are performed mainly based on luminance differences. Note that near iteration number 30, 40, and 50, the merging steps are performed between regions that produce a large increase in perimeter but are very similar in color and that were not merged before due to the high value of the contour term.

Figs. 11 and 12 present an example of BPT above the search partition created with the NWMC and the WEDM criteria, respectively. These images exemplify the case of objects in the scene being correctly gathered in single nodes if the NWMC is used. Using the WEDM criterion, the object information is split among various nodes. To clearly illustrate this point, nodes related to parts of the human faces and their associated subtrees are shown. Note that each face is correctly represented with a single node in the NWMC case (see Fig. 11) whereas in the WEDM case, neither of the two faces is represented with a single node and face regions are split among various nodes (see Fig. 12).

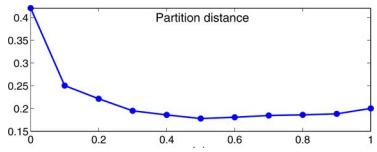

Fig. 14. Mean value of the asymmetric distance computed on the Corel database subset for different α values.

More examples of the improvement achieved with the NWMC are presented in Fig. 13. In this case, we do not present the complete BPT but only those nodes related to the object under study. Row (a) in Fig. 13 shows the original images. In turn, row (b) presents the nodes that represent object regions for the original WEDM criterion [(5)], row (c) those nodes obtained by the criterion based on normalized color but without the contour complexity term (6), and row (d) those nodes obtained by the proposed NWMC criterion (8), respectively. As it can be seen, for this type of semantic objects, the rational behind the definition of the new criterion works perfectly: the use of a normalized color term already improves the quality of the created nodes [row (c)] which is further improved by the inclusion of the contour complexity term [row (d)].

Experiment 5: The behavior illustrated in the previous figures is further analyzed in this experiment by using the Corel database subset. For the whole subset, the node in the BPT leading to the smallest symmetric distance with respect to each manually segmented object has been selected. Statistics of the symmetric distances achieved with the nodes selected from BPTs created using the WSDM, WEDM and NWMC merging criteria are presented in Table IV. Note that, in the case of more generic objects as those represented in the Corel subset (not all of them homogeneous in color), the NWMC also outperforms both the WSDM and the WEDM criteria.

Experiment 6: The next experiment compares the use of the WSDM, WEDM and NWMC for the creation of the search

TABLE IV
SYMMETRIC DISTANCE OVER THE BPT ON THE COREL SUBSET. MEAN AND VARIANCE VALUES ARE MULTIPLIED BY 10^2

Symmetric distance	WSDM	WEDM	NWMC
Mean	24.62	25.65	17,77
σ^2	4.45	4.30	2.55

TABLE V
SYMMETRIC DISTANCE OVER THE BPT ON THE MPEG-7 SUBSET. MEAN AND VARIANCE VALUES ARE MULTIPLIED BY 10^2

Symmetric distance	WSDM	WEDM	NWMC
Mean	44.28	41.87	27.16
σ^2	3.55	3.56	2.75

space of the BPT with the MPEG-7 face database subset (see Section III). As previously, for the whole subset, the node in the BPT leading to the smallest symmetric distance with respect to each manually segmented human face has been selected. Statistics of the symmetric distances achieved with the nodes selected from BPTs created using the WSDM, WEDM and NWMC merging criteria are presented in Table V. As it can be seen, results confirm the improvements introduced by the NWMC criterion. Nevertheless, it has to be noticed that the symmetric distances achieved with the MPEG-7 database are larger (almost the double) than those obtained with the Corel database. This is due to the further difficulty of the MPEG-7 database scenarios, in which all images are complex and several ones presenting illumination problems.

In order to complete the study of the image representation creation, here we analyze the impact of the value of the parameter α in the performance of the NWMC merging criterion.

Experiment 7: In this experiment, the sensitivity of NWMC with respect to the value of the parameter α is analyzed. We use the 160 objects manually segmented from the Corel database subset. The asymmetric distances between the partitions obtained with each of the different α values (using as stopping criterion $N_{\text{reg}} = 50$ regions) and the object partition (ground truth) are computed for each object. Fig. 14 shows the mean value of the asymmetric distance computed over the object database. As it was previously commented, the sensitivity of the segmentation process is very low with respect to variations of the alpha parameter around $\alpha = 0.5$ and, therefore, this value is used.

V. OBJECT DETECTION ON BINARY PARTITION TREES

Nodes in the search space of the BPT represent good markers (anchor points) for object detection algorithms but the accurate analysis of complex images requires going from the object marker to a precise object definition. The accuracy partition ensures that the image representation is precise enough, giving the possibility of refining the object representation. Note that, given the tree structure, the accuracy partition is finer than the search partition (see Sections II and III); that is, all the contours present in the search partition are represented in the accuracy partition as well. This characteristic facilitates the combination of the information from both partitions.

A. Node Extension

The proposed way to combine the information from both partitions for object detection is by making use of the concept of *extended node*. A node in the BPT can be a good estimate of an object, giving information about the position and scale at which the object can be found. However, as the BPT has been built with a generic purpose, nodes in the BPT will very likely not represent complete common objects (e.g., nonhomogenous color objects).

A specific object detection algorithm can make use of the *a priori* information available about the object and improve the object representation yielded by the BPT node.

One possible approach is to use shape information to modify the area of support associated with the position and size pointed by a node. In our work, this is done by fitting a shape model of the object to the node (that is, to the region associated with the node), and then using this shape to modify the area of support of the node, by adding or removing regions from the set of regions that are initially associated with the node.

The shape fitting is performed with a shape matching technique based on distance transforms. Let I be a binary image where zero-valued pixels represent the contour points of a node. A distance transform on I, DT_I is an image where each pixel value denotes the distance to the nearest contour point in I [21], [9]. The matching is performed by transforming the reference shape model by a set of allowed geometrical transformations (typically translation, rotation and scaling) and correlating the transformed template against the DT_I image. The similarity measure between the transformed shape and the node is

$$D(S, N) = \frac{1}{N_S} \sum_{x \in S} \mathrm{DT}_I(x) \tag{11}$$

where S is the contour of the transformed template $\bar{S}$ and N_S is the number of contour points in S. The goal is to search for the best matching; that is, the parameters of the transformation that minimize the similarity measure $D(S, N)$.

The advantage of matching with distance transforms is their ability to handle noisy or imperfect data [25]. In our case, it allows filling gaps in the contour definition due to low gradient values in the image or errors in the segmentation. This ability is illustrated in the examples of Sections V-B and V-C.

Two distance transforms have been used in this work: Chamfer distances [5] and binary distances. Binary distances assign a distance value equal to 1 to a crown (in our work, typically, of 5 pixels width) around the contours, whereas the remaining points in the space receive a distance value equal to 0. In our experiments, achieve better results when the reference shape is a good model of the objects being searched, allowing partial matchings when the node is an incomplete representation of the object. Chamfer distances perform better than binary distances when there is more variability between the shape model and the objects in the image. However, Chamfer distances may lead to local minima when the node is not a correct estimate of the complete object. In our work, the allowed transformations for the reference shape are translation, rotation and scaling. For each node, the set of possible parameter values is bounded and quantized taking into account the node width,

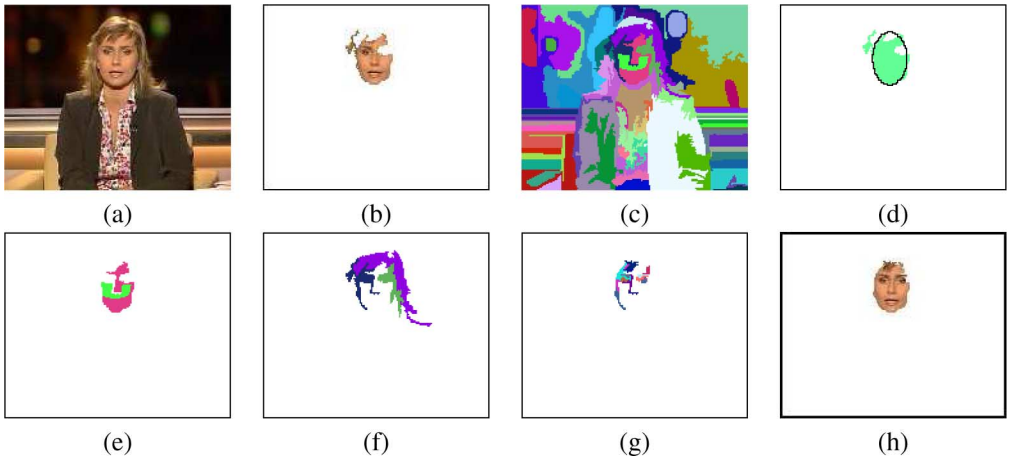

Fig. 15. Node extension: (a) Original image, (b) best representation of the face in the BPT, (c) search partition, (d) shape matching with an elliptical shape model, (e) set of initial included regions from the search partition, (f) set of analyzed regions from the search partition, (g) set of included regions from the accuracy partition, and (e) final extended node.

height and position in the image and an exhaustive search in the parameter space is performed.

The shape matching process is illustrated in Fig. 15 with an example of node extension for face detection. The reference contour which is used as model for the human face shape is an ellipse. The best representation of the face in the BPT is the node shown in Fig. 15(b). Note that there is a region in the node [see the initial partition in Fig. 15(c)] that spans part of the face and the hair, and there is a missing region (the right eye) in the face representation. In spite of that, the reference contour is correctly located around the face. Fig. 15(d) shows the result of the shape matching using a binary distance transform.

The next step in the node extension process is the redefinition of the support area of the node, which can be performed in several ways. Here we propose a simple, generic solution that uses only geometric information. It consists in analyzing the degree of overlap between the regions from the search partition and the fitted shape model $\bar{S}$. Regions in the search partition that are completely included or partially included but that do not extend too far from the shape $\bar{S}$ (typically, if they extend less than 10% of the minor axis of the ellipse) are included in the extended node. Regions that do not overlap $\bar{S}$ are not included. Fig. 15(e) presents the set of regions from the search partition associated with the node in the previous example that are totally included in the shape $\bar{S}$.

For the remaining regions (that is, regions partially included but extending far from the shape $\bar{S}$), the analysis is performed in terms of the accuracy partition, to improve the precision of the representation. Fine partition regions that (do not) overlap with the shape are (removed from) included in the extended node. Fig. 15(f) presents the regions in the previous example that are analyzed using the accuracy partition information. In turn, Fig. 15(g) shows the set of regions from the accuracy partition that are included in the extended node and Fig. 15(h) shows the final area of support of the extended node.

B. Examples

In this subsection, we further illustrate the usefulness of the proposed representation for the detection of objects with different characteristics; namely, human faces, traffic signals, butterflies and cars. Here, complete object detection algorithms are not developed and only simple approaches to exploit the proposed representation are presented. Specifically, we only use shape descriptors of the objects being searched for performing node extension. Note that a real application should use more information about the object to simplify the search and to make the final decision. An example of a complete object detector will be briefly presented in Section V-C.

Two sets of examples are presented in Figs. 16 and in 17. Fig. 16 has five examples dealing with objects whose shapes can be characterized by simple geometric figures: the first two sets of images are related to the detection of human faces and the following three sets of images to the traffic sign case. In turn, Fig. 16 has six examples dealing with objects whose shapes require more complex models: the first four cases analyze the detection of butterflies and the following two cases that of cars.

In all the examples presented in both figures the search partitions were created using the WEDM merging criterion and the AMC stopping criterion with $T_{\mathrm{ACM}} = 0.12$, starting from an accuracy partition with 500 regions. Nodes in the search space of the BPTs were created with the NWMC merging criterion proposed in Section IV-A ($\alpha = 0.5$) which encourages the creation of compact and color-homogeneous nodes. Therefore, in many cases, the objects are correctly represented by a single node in the tree. We only show examples where this does not happen and the extension step is required. Given the type of shapes of the objects to be detected, in the examples of Fig. 16 the shape matching process uses a binary distance whereas in those in Fig. 17, a Chamfer 3–4 distance transform [5] is applied.

In the examples of face detection [Fig. 16(a) and (b)], the shape of a human face is modeled with an ellipse. In the first case, note that the shape model is correctly matched in spite of having a node whose area of support spans the face and the neck. This allows for including a missing region (the right eye) and removing of a very common leakage in human face segmentation (the neck area). In the second case, note that the shape model helps to correctly complete the region of support of the human

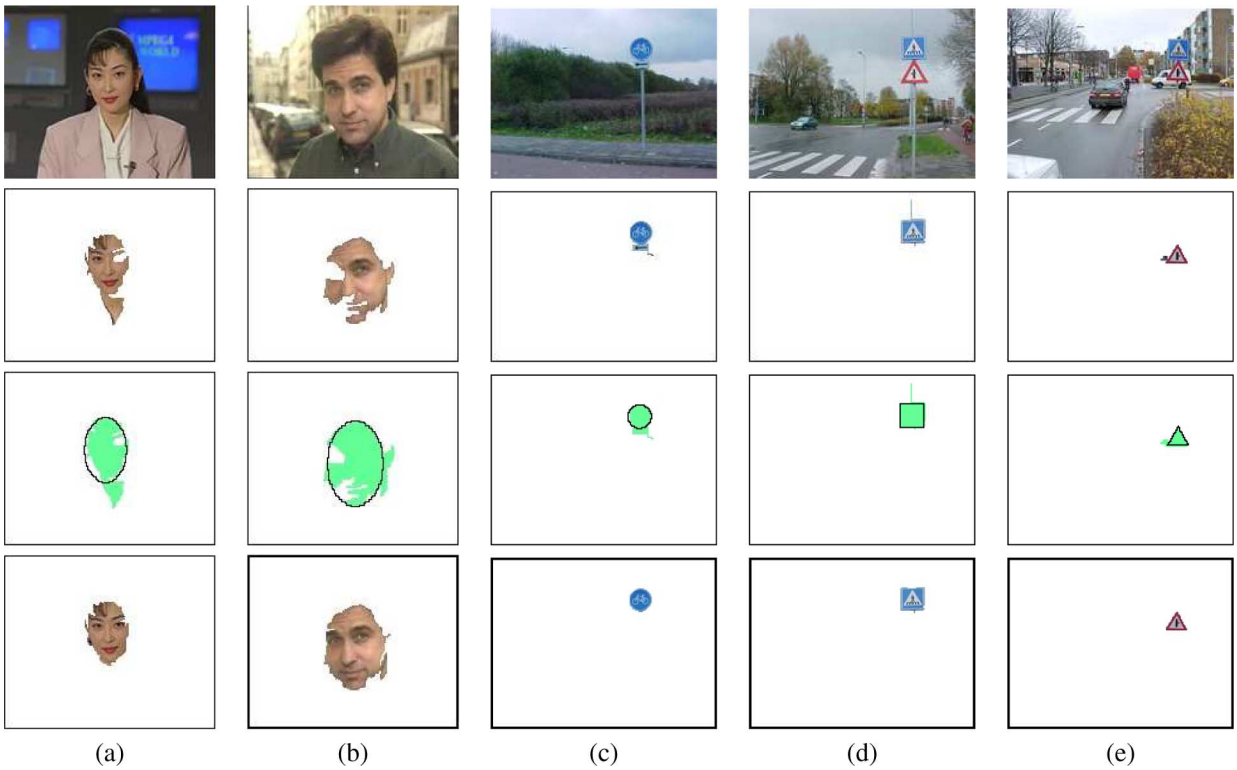

Fig. 16. Improvement in the representation of objects achieved by the extension of the nodes, for the human face and traffic sign detection applications. First row: original images. Second row: object nodes. Third row: shape fitting. Fourth row: extended nodes.

face in spite of not being a perfect frontal view (and, therefore, the ellipse model not perfectly fitting the object).

Examples in Fig. 16(c)–(e), show pictures from a traffic sign image database [14]. For the bike sign (c), the model is a circle; for the pedestrian crossing sign (d), a square; whereas for the junction sign (e), a triangle. In these cases, observe that the image representation is capable to correctly represent, as nodes in the tree, regions being very good markers of small objects (the traffic signs) even in cluttered backgrounds. Moreover, the node extension is able to remove the leakages, that are due to the presence of other objects of similar color close to the traffic signs to be detected.

Six examples of images are presented in Fig. 17 illustrating the use of more complex shape descriptions. The first four examples belong to the *butterfly* class and the last two to the *car* class of the Corel subset. As previously, the image representation has been created following the procedure proposed in this paper. The butterfly and car masks have been manually created using as example, in every case, a different Corel database image of the same class.

Fig. 17(a) presents the example of a wrong representation of the object in the search space (two different objects are merged in the best node, leading to a very different shape from that of the original object) and the successful extension of such a node. Note that, in this case, the shape is correctly matched to a part of the region represented by the node and that the correct matching has required the re-orientation of the model shape. In turn, Fig. 17(b) presents the correct extension of a node in an image in which the background and the object present very similar colors and where shadows are present.

Fig. 17(c) shows an example of the unsuccessful extension of the selected node. This is due to the fact that the model shape being used does not actually correspond to the type of object being sought. The type of butterfly present in the image is better represented by the shape model used in Fig. 17(d) and, in this case, the node is correctly extended leading to a perfect detection of the object.

For the last two sets of images, Fig. 17(e) and (f), a given car shape is used to detect the objects in the scene. Note that the node extension correctly fits the shapes, and mainly acts adding the missing regions to the complete representation, even if they are very different in color as it is the case of the wheels.

In order to assess the usefulness of the proposed region-based representation model, even when using simple object models as those above presented for human faces and traffic signs, two different experiments have been conducted.

Experiment 8: Table VI presents the statistics of the asymmetric (for the accuracy and search partitions) or symmetric (for the BPT and extended BPT nodes) distances between the 116 human faces manually segmented from the MPEG-7 database subset and the selected i) union of regions in the acuracy partition, ii) union of regions in the search partition, iii) node in the search space of the BPT, and iv) extended node in the BPT. As it can be observed, in this database containing mainly images of high complexity, the set of nodes proposed in the search space of the BPT present lower quality than the best union of regions that can be selected from the search partition. The reason for this effect is mainly twofold: First, as previously commented, images in this database present illumination problems and, second, human faces are complex objects (non homogeneous in color).

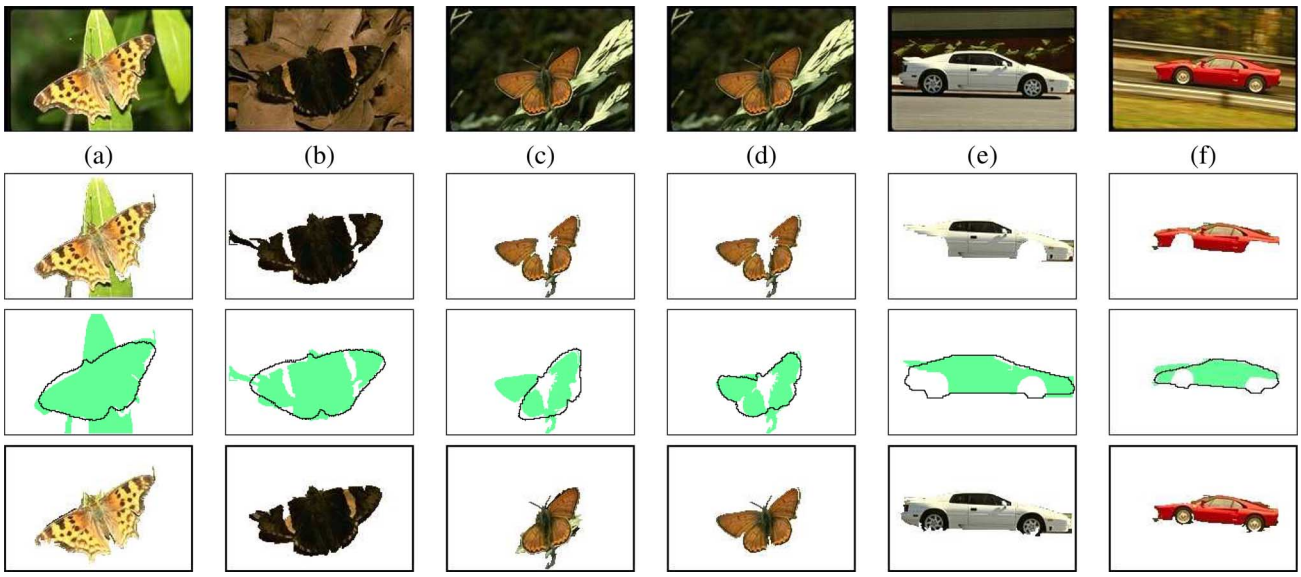

Fig. 17. Improvement in the representation of objects achieved by the extension of the nodes, for the butterfly and car detection applications. First row: original images. Second row: object nodes. Third row: shape fitting. Fourth row: extended nodes.

TABLE VI
ASYMMETRIC DISTANCE FOR THE SELECTED UNION OF REGIONS FROM THE ACCURACY PARTITION AND FROM THE SEARCH PARTITION, AS WELL AS SYMMETRIC DISTANCE FOR THE SELECTED BPT NODE AND FOR THE SELECTED EXTENDED NODE WITH RESPECT TO THE MPEG-7 SUBSET. MEAN AND VARIANCE VALUES ARE MULTIPLIED BY 10^2

	Asymmetric Distance		Symmetric Distance	
	Accuracy Part.	**Search Part.**	**BPT Node**	**Ext. Node**
Mean	5.99	16.48	28.18	20.08
σ^2	0.07	0.89	3.05	1.52

TABLE VII
ASYMMETRIC DISTANCE FOR THE SELECTED UNION OF REGIONS FROM THE ACCURACY PARTITION AND FROM THE SEARCH PARTITION, AS WELL AS SYMMETRIC DISTANCE FOR THE SELECTED BPT NODE AND FOR THE SELECTED EXTENDED NODE WITH RESPECT TO THE MPEG-7 SUBSET. MEAN AND VARIANCE VALUES ARE MULTIPLIED BY 10^2

	Asymmetric Distance		Symmetric Distance	
	Accuracy Part.	**Search Part.**	**BPT Node**	**Ext. Node**
Mean	6.31	22.25	22.60	16.29
σ^2	0.97	3.52	3.58	1.71

TABLE VIII
BPT FILTERING RESULTS: PERCENTAGE OF NODES THAT ARE FILTERED OUT BY EACH CRITERION

	Act	**No Act**	**Color**	**Size**	**AR**
MPEG	16.03	87.97	45.17	52.58	47.34

TABLE IX
DETECTION AND SEGMENTATION RESULTS

	Percentage				Symmetric Distance		
	Det	**DE**	**Seg**	**SE**	**AccD**	**AccS**	**AccSE**
MPEG	84.8	15.2	89.5	10.5	0.73	0.82	0.54

Therefore, nodes in the BPT cannot in some cases represent a perfect marker of the object in the scene. This problem is further discussed in Section V-C.

Experiment 9: Table VII presents the statistics of the asymmetric (for the accuracy and search partitions) or symmetric (for the BPT and extended BPT nodes) distances between the 45 traffic signals manually segmented from the related database [14] and the selected i) union of regions in the accuracy partition, ii) union of regions in the search partition, iii) node in the search space of the BPT, and iv) extended node in the BPT. As it can be observed, in this database which contains images of medium/high complexity [see columns (c), (d), and (e) in Fig. 16], the nodes proposed by the BPT are very close in performance to the complete set of possible unions of regions proposed by the search partition. Moreover, the use of the extended node leads to a decrease on the symmetric distance; that is, it allows improving the object representation. Finally, it has to be commented that the statistical behavior is very similar for the three types of signals in the database (slightly better for the triangular signal).

C. Object Detection Strategy. Example: Face Detection

In this subsection, we illustrate the usefulness of the proposed representation for the detection of a specific class of objects: human faces. Although proposing complete object detection algorithms is out of the scope of this paper, for the sake of completeness we briefly present here the type of methodology that can be used relying on the BPT representation and the concept of node extension. For a complete description, the reader is referred to [30].

As it has been stated in Section I, object detection requires both an image representation suitable for this task and a useful object model. The object model is, in pattern recognition terms, the training or design cycle of a classifier [10]. It comprises the selection of a set of representative features, the proposal of a classifier model and the training and testing of the classifier. Concerning the face model, it is built as a combination of one-class classifiers [29] based on visual attributes which are

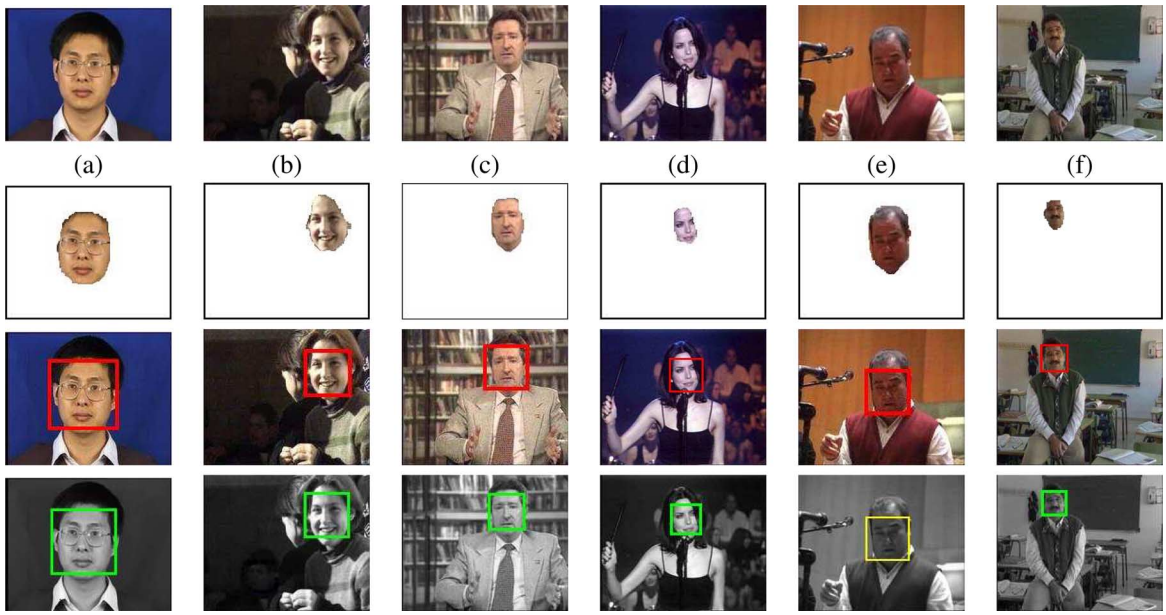

Fig. 18. Examples of the (a) XM2VTS [17], (b) to (f) MPEG-7 databases and associated results. Second row: joint detection and segmentation using our technique [30]. Third row: detection using Viola and Jones [33] (OpenCV implementation). Forth row: detection using Schneiderman [27], [28] (online face detector demo).

measured on the regions by a set of descriptors. Each descriptor is therefore associated with a simple one-class classifier which is typically trained using sample data.

Taking into account the way we use the descriptors for the selection of candidates and for the final classification, we organize them into three groups: i) generic descriptors for the simplification of the search space (*BPT filtering*), ii) shape descriptors for the accurate selection of candidates (*node extension*), and iii) specific descriptors for the final object detection (*classification*).

BPT filtering aims to discard as many nodes as possible from the search space of the BPT by a fast analysis of the regions associated with these nodes. The goal is to analyze the patterns using a set of generic descriptors computed for each node, removing those nodes that significantly differ from the characterization proposed by the face model. Generic descriptors are descriptors associated with low-level (basic) visual attributes. They are simple and relatively easy to measure; that is, with a low computational cost. We work with the following generic descriptors: Color mean, Aspect ratio, Oriented aspect ratio, Size, Orientation and Position.

Node extension, as explained in Section V-A, is used to accurately redefine the area of support of the remaining nodes, if necessary. For the case of face detection, a simple shape description based on modeling the face as an ellipse is sufficient. We work with an elliptical model that describes the general shape of a face, and allows adapting the area of support of the remaining BPT nodes to the shape of the object being searched.

Classification evaluates the set of specific descriptors on each extended node candidate and assigns them a face or no-face label. Each specific descriptor is associated with a simple classifier that outputs, for each candidate, its likelihood of being a face instance. The outputs are combined into a global probability or face likelihood. Finally, the most likely candidates are selected as face instances. Specific descriptors are descriptors related to attributes that are specific to faces. Generally, the selection of these attributes implies further knowledge about faces. They are usually more complex and costly to compute than generic descriptors but they are evaluated on a very reduced set of regions. We work with the following set of specific descriptors: Dominant color, Symmetry, Hausdorff distance, PCA distance in the feature space and PCA distance from the feature space. For the experiments presented in this work, descriptors are combined, for simplicity, by a product combination of estimated likelihoods.

In Table VIII, the results of the filtering step on the MPEG-7 subset database are presented, showing the percentages of active (Act) and nonactive (No Act) nodes after the filtering step. As it can be seen, the use of generic descriptors largely reduces the amount of nodes to be further analyzed. Moreover, the percentage of nodes that are filtered by the different criteria are presented. Note that a node can be filtered by more than one criterion, as it is typically the case for the color, size and aspect ratio criteria. Finally, the filtering effect of some of the above proposed generic descriptors is not reported in this table since, for this specific application, they are not relevant (e.g.,: Orientation or Position).

Table IX presents the final detection and segmentation results on the MPEG-7 subset database. Faces are said to be correctly detected (Det) if, (i) in the case of underestimating the region, the extended node contains at least one eye and the mouth of the person and (ii) in the case of overestimating the region, the center of mass of the extended node is inside the face area. The correctly detected faces are, afterwards, classified into correctly segmented (Seg) or producing segmentation errors (SE). Again, these errors can be due to an underestimation (lack of parts of the face) or an overestimation (inclusion of parts of the background) in the extended node. Finally, figures of the segmentation accuracy are reported; that is, the symmetric distance between the extended node and the related ground truth object partitions are computed for all the detected faces (AccD), for the correctly

segmented (AccS) and for those that produce segmentation errors (AccSE).

Examples and results of three different databases are presented in Fig. 18. The various examples have been selected to cover various aspects of the system behavior: people wearing glasses, bad illumination scenarios, cluttered background, deviations from the frontal face pose, etc.

Fig. 18 also shows detection results for two state-of-the-art face detection techniques; namely, Viola and Jones [33] and Schneiderman [27] systems. These are two very robust and fast cascade methods which use simple and fast evaluation features in early stages of the cascade and more complex and discriminative features at later stages. The output of these systems is a rectangle around the main facial features, which may contain part of the background or may lack some skin areas. As it can be seen in Fig. 18, regarding the final detected faces the three algorithms yield high quality results, being the proposed one much more accurate in terms of face segmentation, given its region-based nature.

Systems such as [33] and [27] imply a costly training phase but perform extremely fast. In the training phase, they require a large amount of face and no-face samples since later stages of the cascade are trained with false positives output by simpler classifiers at early stages. Nevertheless, in the execution phase, they present real time performance (under the conditions established in [33], an image of size 384×288 is processed in 67 ms on a Pentium III 700 MHz; whereas in [27], using a 32×24 input window, the detector evaluates a 300×200 image in under 180 ms on an Athalon 1.8 GHz processor). Although there is a large number of rectangles to evaluate, strategies such as the use of the integral image representation [33] and the feature centric evaluation [27] combined with the use of the cascade make these algorithms very fast. Regarding the proposed region-based approach, its training phase is much simpler since it implies a relatively reduced set of face images for the texture classifiers (around 1000 face images suffices), and only positive samples are needed (the system is composed of one-class classifiers). In the execution phase, the proposed technique does not reach real time performance, since the BPT construction is costly (on average, for an image of size 176×144, this step takes 1.03 seconds on a Pentium-M 1.87 GHz).

Nevertheless, it has to be highlighted that, as demonstrated in this paper, the building of the image representation is a generic step in object detection. Therefore, the image representation can be adopted for subsequent detection of different object classes. This concept of reusability cannot be applied to systems such as [33] and [27], since they would require a complex training process for each new object to be detected.

VI. Conclusion

This paper has discussed the use of Binary Partition Trees (BPT) for object detection. BPT are hierarchical region-based representations of images. They define a reduced set of regions that covers the image support and that spans various levels of resolution. In this paper, several issues related to the use of BPT for object detection have been studied: Concerning the tree construction, we have analyzed the compromise between computational complexity reduction and accuracy. This led us to define two parts in the BPT: one providing accuracy (the *BPT accuracy space*) and one representing the search space for the object detection task (the *BPT search space*). Then we have objectively compared various similarity measures for the tree construction. We concluded that different similarity criteria should be used for the two parts of the BPT: The Weighted Euclidean Distance between regions Model (WEDM) may be used for the definition of the *BPT accuracy area* and the Normalized Weighted Euclidean distance between Models with Contour complexity (NWMC) for the *BPT search space*. The transition between the accuracy and the search spaces in the BPT is defined by the so-called *Accumulative Merging Cost* (AMC). Finally, we have discussed a generic object detection strategy based on BPT. The notion of node extension was proposed and discussed. Several object detection examples illustrating the generality of the approach and its efficiency were reported.

References

[1] T. Adamek, "Using Contour Information and Segmentation for Object Registration, Modeling and Retrieval," Ph.D. dissertation, Dublin City Univ., Dublin, Ireland, 2006.

[2] S. Agarwal, A. Awan, and D. Roth, "Learning to detect objects in images via a sparse, Part-based representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 11, pp. 1475–1490, Nov. 2004.

[3] Y. Amit, *2D Object Detection and Recognition, Models, Algorithms and Networks*. Cambridge, U.K.: MIT Press, 2002.

[4] C. Ballester, V. Caselles, and P. Monasse, *The Tree of Shapes of an Image, ESAIM: COCV*, vol. 9, pp. 1–18, Jan. 2003.

[5] G. Borgefors, "Hierarchical chamfer matching: A parametric edge matching algorith," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 10, no. 11, pp. 425–429, Nov. 1988.

[6] P. Burt and E. Adelson, "The laplacian pyramid as a compact image code," *IEEE Trans. Commun.*, vol. COM-31, no. 4, pp. 532–540, Apr. 1983.

[7] J. Cardoso and L. Corte-Real, "Toward a generic evaluation of image segmentation," *IEEE Trans. Image Process.*, vol. 14, no. 11, pp. 1773–1782, Nov. 2005.

[8] P. Correa and F. Pereira, "Stand-alone objective segmentation quality evaluation," *EURASIP J. Appl. Signal Process.*, vol. 2002, no. 4, pp. 389–400, 2002.

[9] O. Cuisenaire and B. Macq, "Fast euclidean distance transformation by propagation using multiple neighborhoods," *Comput. Vis. Image Understand.*, vol. 76, no. 2, pp. 163–172, Nov. 1999.

[10] R. O. Duda, P. Hart, and D. Stork, *Pattern Classification*, 2nd ed. New York: Wiley, 2001.

[11] C. Ferran and J. R. Casas, "Object representation using colour, Shape and structure criteria in a binary partition tree," in *Proc. Int. Conf. Image Processing*, Genova, Italy, 2005, vol. III, pp. 1144–1147.

[12] L. Garrido, "Hierarchical Region Based Processing of Images and Video Sequences: Application to Filtering, Segmentation and Information Retrieval," Ph.D. dissertation, Techn. Univ. Catalonia, Spain, 2002.

[13] L. Garrido and P. Salembier, "Region based analysis of video sequences with a general merging algorithm," in *Proc. IX European Signal Processing Conf. EUSIPCO*, Rhodes, Greece, Sep. 1998, vol. III, pp. 1693–1696.

[14] C. Grigorescu and N. Petkov, "Distance sets for shape filters and shape recognition," *IEEE Trans. Image Process.*, vol. 12, no. 10, pp. 1274–1286, Oct. 2003.

[15] L. Huihai, J. C. Woods, and M. Ghanbari, "Binary partition tree analysis based on region evolution and its application to tree simplification," *IEEE Trans. Image Process.*, vol. 16, no. 4, pp. 1131–1138, Apr. 2007.

[16] L. Huihai, J. C. Woods, and M. Ghanbari, "Binary partition tree for semantic object extraction and image segmentation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 17, no. 3, pp. 378–383, Mar. 2007.

[17] K. Messer *et al.*, "XM2VTSbd: The extended M2VTS database," presented at the 2nd Conf. Audio and Video Based Biometric Personal Verification, New York, 1999.

[18] P. Monasse and F. Guichard, "Fast computation of a contrast-invariant image representation," *IEEE Trans. Image Process.*, vol. 5, no. 9, pp. 860–872, May 2000.
[19] MPEG, Description of MPEG-7 Content Set, ISO/IEC JTC1/SC29/WG11/N2467. Atlantic City, NJ, Oct. 1998.
[20] M. Pietikainen and A. Rosenfeld, "Image segmentation by texture using pyramid node linking," *IEEE Trans. Syst., Mach., Cybern.*, vol. SMC-11, no. 12, pp. 822–825, Dec. 1981.
[21] A. Rosenfeld and J. Pfaltz, "Distance functions in digital pictures," *Pattern Recognit.*, vol. 1, no. 1, pp. 33–61, Jul. 1968.
[22] P. Salembier and L. Garrido, "Binary partition tree as an efficient representation for image processing, segmentation and information retrieval," *IEEE Trans. Image Process.*, vol. 9, no. 4, pp. 561–575, Apr. 2000.
[23] P. Salembier and F. Marques, "Region-based representation of image and video: Segmentation tools for multimedia services," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 9, no. 8, pp. 1147–1167, Dec. 1999.
[24] P. Salembier, A. Oliveras, and L. Garrido, "Anti-extensive connected operators for image and sequence processing," *IEEE Trans. Image Process.*, vol. 7, no. 4, pp. 555–570, Apr. 1998.
[25] O. Salerno, M. Pardas, V. Vilaplana, and F. Marques, "Object recognition based on binary partition trees," in *Proc. IEEE Int. Conf. Image Processing*, Singapore, Oct. 2004, pp. 929–932.
[26] H. Samet, *The Design and Analysis of Spatial Data Structures*. Reading, MA: Addison Wesley, 1990.
[27] H. Schneiderman, "Feature-centric evaluation for efficient cascaded object detection," presented at the IEEE Conf. Computer Vision and Pattern Recognition, 2004.
[28] H. Schneiderman, "Learning a restricted bayesian network for object detection," presented at the IEEE Conf. Computer Vision and Pattern Recognition, 2004.
[29] D. Tax, "One-Class Classification: Concept Learning in the Absence of Counter-Examples," Ph.D. dissertation, Tech. Univ. Delft, Delft, The Netherlands, 2001.
[30] V. Vilaplana and F. Marques, "Face detection and segmentation on a hierarchical image representation," presented at the 15th Eur. Signal Processing Conf. EUSIPCO, Poznand, Poland, Sep. 2007.
[31] P. Villegas and X. Marichal, "Perceptually-weighted evaluation criteria for segmentation masks in video sequences," *IEEE Trans. Image Process.*, vol. 13, no. 8, pp. 1092–1103, Aug. 2004.
[32] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," presented at the IEEE Conf. Computer Vision and Pattern Recognition, Dec. 2001.
[33] P. Viola and M. Jones, "Robust real-time object detection," *Int. J. Comput. Vis.*, vol. 57, no. 2, pp. 137–154, 2004.
[34] T. Vlachos and A. G. Constantinides, "Graph-theoretic approach to color picture segmentation and contour classification," *IEE Proc.*, vol. 130, no. 1, pp. 36–45, Feb. 1993.
[35] M. Yang, D. Kriegman, and N. Ahuja, "Detecting faces in images: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 1, pp. 34–58, Jan. 2002.

Veronica Vilaplana (M'97) received the degree in mathematics and the degree in computer sciences from the University of Buenos Aires, Argentina, in 1993 and 1994, respectively. She is currently pursuing the Ph.D. degree in the Department of Signal Theory and Communications, Technical University of Catalonia (UPC), Spain.

In 2002, she joined the Department of Signal Theory and Communications, UPC. She is Assistant Professor lecturing in the area of digital signal and image processing. Her research interests include image modeling and representation, object detection and tracking, pattern recognition, and biometric systems.

Ferran Marqués (M'91) received the degree in electrical engineering and the Ph.D. degree from the Technical University of Catalonia (UPC), Barcelona, Spain, in 1988 and 1992, respectively.

From April 1989 to June 1990, he was with the Swiss Federal Institute of Technology, Lausanne (EPFL), Switzerland, and in 1991, he was with the Signal and Image Processing Institute, University of Southern California, Los Angeles. He became an Associate Professor in 1995. Since 2003, he has been a Full Professor with the TSC Department, UPC, where he is lecturing in the area of digital signal and image processing. He served as Associate Dean for International Relations of the Telecommunication School (ETSETB) at UPC (1997–2000) and as President of the European Association for Signal Processing EURASIP (2002–2004). He is the author or coauthor of more than 100 publications that have appeared as journal papers and proceeding articles, five book chapters, and four international patents.

Dr. Marqués served as the Associate Editor of the *Journal of Electronic Imaging* (SPIE) in the area of Image Communications (1996–2000), as member of the EURASIP *Journal of Applied Signal Processing* Editorial Board (2001–2003), and, since 2006, he has been a member of the EURASIP *International Journal of Image and Video Processing* Editorial Board. He has acted as the Guest Editor for the journal *Signal Processing: Image Communication* (Elsevier) (Special Issue on Image Processing for 3-D Imaging) and for the EURASIP *Journal of Applied Signal Processing* (Special Issue on Signal Processing for 3D Imaging and Virtual Reality). He has been Co-Chairman of the EUROIMAGE International Conference on Augmented Virtual Enviroments and 3-Dimensional Imaging (ICAV3D-2001, Mykonos, Greece, May 2001), Special Sessions Co-Chairman in the International Conference on Image Processing (ICIP-2003, Barcelona, Spain, September 2003), and Technical Chairman of the 4th International Workshop on Content-Based Multimedia Indexing (CBMI-2005, Riga, Latvia, June 2005). He won the Spanish Best Ph.D. Thesis in Electrical Engineering Award in 1992.

Philippe Salembier (M'95) received the degree from the Ecole Polytechnique, Paris, France, in 1983, the degree from the Ecole Nationale Supérieure des Télécommunications, Paris, in 1985, and the Ph.D. degree from the Swiss Federal Institute of Technology, Lausanne (EPFL), in 1991.

He was a Postdoctoral Fellow at the Harvard Robotics Laboratory, Cambridge, MA, in 1991. From 1985 to 1989, he worked at Laboratoires d'Electronique Philips, Limeil-Brevannes, France, in the fields of digital communications and signal processing for HDTV. In 1989, he joined the Signal Processing Laboratory of the EPFL to work on image processing. At the end of 1991, after a stay at the Harvard Robotics Laboratory, he joined the Technical University of Catalonia, Barcelona, Spain, where he is currently a Professor lecturing on the area of digital signal and image processing. His current research interests include image and sequence coding, compression and indexing, image modeling, segmentation, video sequence analysis, mathematical morphology, level sets, and nonlinear filtering. In terms of standardization activities, he has been particularly involved in the definition of the MPEG-7 standard ("Multimedia Content Description Interface") as chair of the "Multimedia Description Scheme" group between 1999 and 2001.

Dr. Salembier served as an Area Editor of the *Journal of Visual Communication and Image Representation* (Academic Press) from 1995 until 1998 and as an AdCom officer of the European Association for Signal Processing (EURASIP) from 1994 until 1999. He has edited (as guest editor) special issues of *Signal Processing* on "Mathematical Morphology" (1994) and on "Video sequence analysis" (1998). He has also co-edited (with Prof. F. Pereira) a special issue of *Signal processing: Image Communication* on MPEG-7 Technology (2000). He was co-Editor-in-Chief of *Signal Processing* between 2001 and 2002. He was a member of the Image and Multidimensional Signal Processing Technical Committee of the IEEE Signal Processing Society between 2000–2006 and was technical chair (with Prof. E. Delp) of the IEEE Int. Conference on Image Processing 2003, organized in Barcelona. Finally, he has served as an Associate Editor of the IEEE TRANSACTIONS ON IMAGE PROCESSING since 2002 and the IEEE SIGNAL PROCESSING LETTERS since 2005.