# HIERARCHICAL REPRESENTATION OF SCENES USING ACTIVITY INFORMATION

*José Luis Landabaso‡, Montse Pardàs‡, Li-Qun Xu†*

‡ Technical University of Catalunya (UPC) † BT Research & Venturing, UK

**ABSTRACT**

Object segmentation and tracking are two key issues in the analysis of scenes for video surveillance or scene understanding applications. This paper addresses the object segmentation task by presenting a new algorithmic contribution in these applications' context. The proposed method combines an adaptive background learning technique with a hierarchical segmentation method based on Binary Partition Trees. The result is a region-based dynamic scene description, where each active region is characterized by a temporal feature, reflecting on the time it remains in the same position of the scene. This description is then used to classify the background and foreground objects of the scene and can also be used as an additional feature for region tracking and scene understanding.

## 1. INTRODUCTION

The development of accurate and robust segmentation and tracking techniques for multiple moving objects in dynamic visual scene analysis is a very challenging issue. It is particularly desirable in the video surveillance field where an automated system allows fast and efficient access to unforeseen events that need to be attended by security guards or law enforcement officers. It also enables tagging and indexing interesting scene activities / statistics in a video database for future retrieval on demand. In addition, such systems are the building blocks of higher-level intelligent vision-based or assisted information analysis and management systems with a view to understanding the complex actions, interactions, and abnormal behaviors of objects in the scene.

Vision-based surveillance systems can be classified in several different ways, depending on the environment in which they are designed to operate. In this paper our focus is on processing videos captured by a single fixed camera overlooking areas in both indoor and outdoor visual monitoring scenarios.

Most of the research activities oriented to segment these scenes adopt the 'background subtraction' as a common approach to detecting foreground moving pixels, whereby the background scene structures are modeled pixel-wise by various statistically-based learning techniques on features such as intensities, colors, edges, texture, etc. The models employed include parallel uni-modal Gaussians [1], mixture of Gaussians [2], nonparametric Kernel density estimation [3], or simply temporal median filtering [4]. A connected component analysis (CCA) is then followed to cluster and label the foreground pixels into meaningful object blobs, from which some inherent appearance and motion features can be extracted. Finally, there is a blob-based tracking process aiming to find persistent blob correspondences between consecutive frames. A complete system based on these principles was presented in [5]. In this paper we focus on a new scheme for the segmentation part which includes the extraction of a different kind of information that can be used both for the blob tracking process and for further processing oriented to scene understanding.

The segmentation procedure that we propose combines an adaptive background learning method based on a mixture of Gaussians as in [2] with a region oriented hierarchical segmentation method. This pixel-based background learning method is chosen because it is robust to illumination changes and will allow creating a 'pixel persistence map' (PPM) indicating, for every pixel, the time elapsed since its color features have changed. The PPM will contain relevant information about the time that the objects in the scene have remained in the same position of the scene. After the background learning and map creation, instead of taking a pixel-based decision as in [2] or [5], a hierarchical segmentation method using Binary Partition Trees (BPT) is applied. Binary Partition Trees [6] provide a structured representation of the regions that can be obtained from an image. The spatial and temporal information are used to create this structure: color information to create an initial partition from which the BPT is initialized, and the pixel persistence information to create the merges that leads to the final BPT. The regions obtained in the nodes of the BPT have a hierarchical ordering corresponding to a new feature: The time that the object has remained in the same position of the scene. This information is stored in the Binary Partition Tree and can be used to decide if an object is included in the background or not. The following tracking process will also use this new feature to establish the blobs correspondence. Furthermore, this information can be very relevant for scene activity estimation in a scene understanding context. One of the advantages of the proposed scheme over the state of the art techniques is the

possibility of distinguishing between the 'old' and 'recent' background. For instance, a car that has been parked (see Figure 3), a bag that has been abandoned (Figure 4), or a person who has entered and then stands still, will not be integrated in the background as is the case with most common background learning techniques. Instead, they are identified as independent objects characterized by the time that they have remained in the same position. Fig. 1 depicts schematically the block diagram of the proposed scheme.

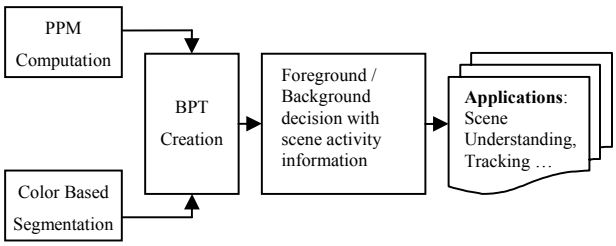



**Fig. 1.** The system block diagram showing the chain of functional modules

The paper is structured as follows. In the next section the techniques for pixel-domain analysis leading to the pixel persistence information are described. Section 3 is devoted to the BPT definition and its creation process. Section 4 illustrates the results obtained with this system. And, finally the paper concludes in Section 5.

## 2. PIXEL PERSISTENCE MAP (PPM)

In video sequences captured with a fixed camera, every pixel usually shows similar values over the time. This is normally true for background pixels except when there are illumination changes. In these 'pixel activity' situations (foreground pixels), the new pixel values usually form a connected region with 'temporal persistence' homogeneity. Imagine for instance an object placed over the floor at a certain moment (see Figure 4). Even if this object has different colors, all the pixels within the object will share the fact that they have been observed in the same place, during the same amount of time.

In order to segment the regions with the criterion of 'temporal persistence' homogeneity, it is crucial to study first the pixel value occurrences along the time. Pixel models are a useful way to rationalize this information without having to accumulate the pixel values until the current frame.

### 2.1. Pixel Model

A probabilistic model is used for every pixel in the image, to account for the recent history of photometric variations of the pixel in RGB color space. We choose a Mixture of Gaussians model $\eta(\mathbf{X}_t, \mu_{i,t}, \Sigma_{i,t})$, with i=1...K, in which each of the Gaussians characterizes a different color appearance over the time. When a certain RGB value, not seen before, appears in a pixel, a new Gaussian is created. This Gaussian characterizes this color appearance from that instant on. The weight associated with each Gaussian will be increased / decreased depending on how often a similar color appears:

$$\begin{aligned} w_{k,t} &= (1-\alpha)w_{k,t-1} + \alpha; \quad \text{if matched} \\ w_{k,t} &= (1-\alpha)w_{k,t-1}; \quad \text{if didn't match} \end{aligned} \tag{1}$$

$1/\alpha$ defines the 'time constant' reflecting the speed at which the weight changes. $w_k$ is a low-pass filtered average of the number of occasions that the kth Gaussian has characterized a color appearance thus far.

The mean and variance of each Gaussian are also updated to allow a model to adapt to slow illumination changes [2][5].

### 2.2. Map Creation

The PPM is built using the weights of the Gaussians which characterize the color values of the current frame. Therefore, the map will assume higher values in areas, e.g. background, where similar colors have appeared frequently and consecutively over the recent history. On the contrary, the map will show lower values in areas where new colors, e.g. due to a moving object, have appeared that sustain for a shorter period of time.

### 2.3. Temporal Persistence

Although the weight associated to each Gaussian is enough for the segmentation purpose, there exists additional useful information hidden in (1). Effectively, if we solve the recursion in the equation, we can obtain the number of frames elapsed since a Gaussian with initial weight $w_o$ has repeatedly characterized the same color until $w_t = w_f$. Therefore, we can determine for how long an object has had its presence in a certain position:

$$\begin{aligned} & w_t = (1-\alpha)w_{t-1} + \alpha = \\ & (1-\alpha)^t w_o + \alpha \sum_{i=0}^{t-1} (1-\alpha)^i \underset{0<(1-\alpha)<1}{=} (w_o - 1)(1-\alpha)^t + 1 \end{aligned} \tag{2}$$

And, thus:

$$t = \log_{1-\alpha} \frac{(w_t - 1)}{(w_0 - 1)} \tag{3}$$

## 3. BINARY PARTITION TREE (BPT)

A Binary Partition Tree [6] is a structured and compact representation of the regions that can be obtained from an initial partition of an image. Several approaches can be used to create this tree. We have used a segmentation that follows a bottom-up approach. The algorithm first

constructs the Region Adjacency Graph of an initial partition. Using a region-based segmentation, the BPT is then created by keeping track of the regions that are merged at each iteration until one region is obtained. That is, for each pair of neighboring regions a homogeneity measure is assessed, and the pair whose distance is the lowest is merged. The process is iterated until one final region is obtained. An example is shown in Figure 2. In order to show the correspondence of the nodes with the image regions, we have taken an initial partition of only 10 regions in Figure 2 (a). The leaves of the tree in Figure 2 (b) represent the regions of this initial partition. The remaining nodes of the tree represent regions that are obtained by merging the regions represented by its two child nodes. The root node represents the entire image support.

The BPT should be created in such a way that the most meaningful regions are represented in its nodes. In our case, we aim at detecting foreground objects. For this reason we generate an initial partition by merging of flat zones with a spatial color similarity criterion, and from this initial partition, we then create the BPT using the temporal persistence map.

### 3.1. Initial Partition

For the initial partition, a region growing approach [6] is used, which is based on a weighted Euclidean distance, shown in (4), in the YUV color space, where more emphasis is given to the luminance component:

$$d(r_1,r_2)=\sqrt{\gamma\left(\overline{y}_1-\overline{y}_2\right)^2+\frac{(1-\gamma)}{2}\left(\left(\overline{u}_1-\overline{u}_2\right)^2+\left(\overline{v}_1-\overline{v}_2\right)^2\right)} \qquad (4)$$

Regions (r1, r2) are compared using the mean values of their YUV components. Regions are merged until a termination criterion is reached (usually the number of regions or the PSNR obtained when representing the original image by the partition with all regions filled with their mean values).

### 3.2. Construction of the BPT

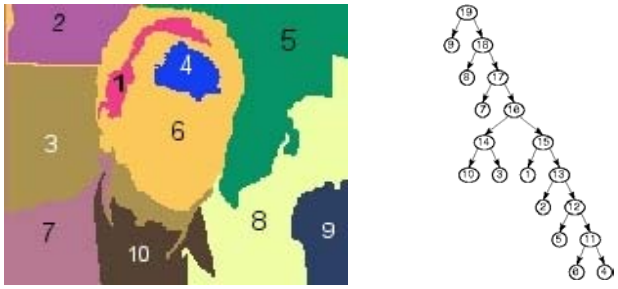
The homogeneity feature that we used to construct the BPT from the initial partition is the PPM. The $L_1$ distance between two regions (r1, r2) is thus defined as:

$$d(r_1,r_2)=\left|\overline{w}_1-\overline{w}_2\right| \qquad (5)$$

That is, the merging order among the regions is defined using the distance between the mean values of the persistence values within the corresponding regions.

In this case, the complete BPT is constructed defining the mergings until a single node is reached.

Using this homogeneity criterion the nodes of the tree are characterized by the time the corresponding object has remained in the same position of the scene. The node with the lowest persistence value corresponds to the background. Upper levels of the tree contain the nodes corresponding to foreground objects.



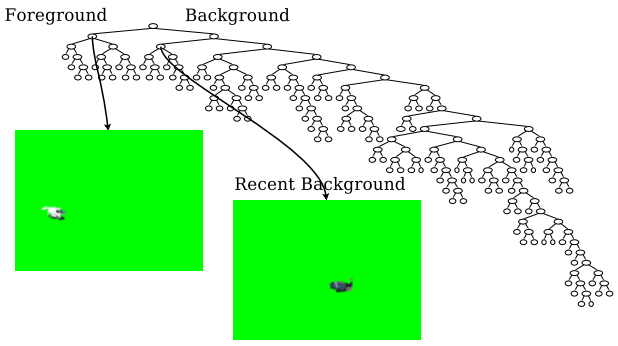

(a) (b)

**Fig. 2.** An example of the initial partition (a) and the Binary Partition Tree created from this initial partition (b).

## 4. RESULTS

The system has been evaluated on standard test sequences such as the set of benchmarking images sequences provided by PETS'2001 and a range of our own captured image sequences (at $768\times576$ pixels) under various compression formats.

(a) Current Frame (b) PPM

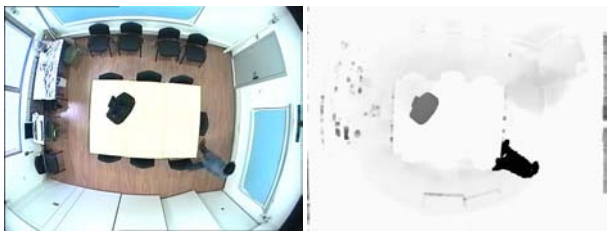



(c) BPT

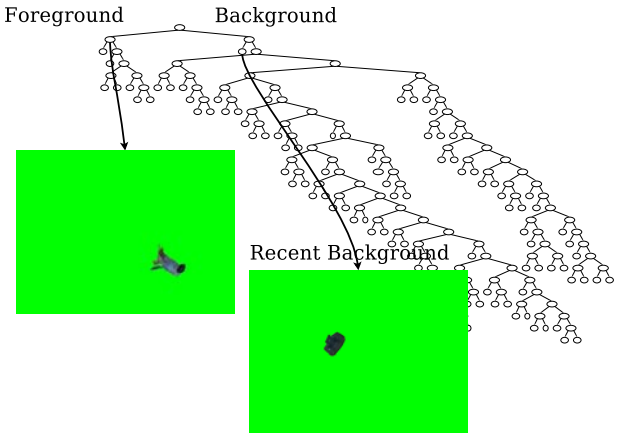
**Fig. 3.** An outdoor scene example of the BPT creation.

Figure 3 (a) shows an example where the green car on the street bend has been parked a few moments (frames) ago, and the white van is currently moving. First, the current frame is segmented using color homogeneity until 500 regions are left. Then, the PPM (b) is used as the

merging criteria to obtain the BPT, in (c). As there is only one foreground object, a second level node contains the entire background. When there are multiple foreground objects, the background scene is represented in lower-level nodes. A termination criterion can be set so that the node representing the background doesn't merge with the nodes containing foreground appearances. Moreover, we can determine using (3), the temporal persistence of the node that has been labeled as 'Recent Background' (Figure 4 (c)). In this case, $t = 169.6$, with $\alpha = 0.02$, and $w_0 = 0$, which indicates very accurately during how many frames the car has been parked (the image shown corresponds to frame 750, CAMERA1 of PETS'01 sequences)



(a) Current Frame (b) PPM




(c) BPT

**Fig. 4.** An indoors scene example of the BPT creation.

Figure 4 depicts another example: in this case, a bag has been left on a table and a person is walking. Similar considerations to the previous case apply here. This situation, though, offers an example of a typical surveillance situation where a suspicious package is detected, and it is urgent to know in which instant it was placed there, so that security agents can easily inspect the recorded images at the exact instant. In this case, $t = 116.5$, with $\alpha = 0.005$, and $w_0 = 0$.

## 5. CONCLUSION

In this paper we have presented a novel approach to object segmentation in video sequences - a hierarchical segmentation procedure using BPTs - which builds upon the temporal information derived from a pixel-wise background learning technique based on mixtures of Gaussians. The advantages of this scheme are the following:

- The foreground / background decisions are taken on a region basis instead of a pixel-basis. The decision is thus more robust to noise effects and does not require a connected component analysis to classify the different foreground objects.
- The detected regions are characterized by the time elapsed since they reached the current position.
- Using the BPT we can separate the foreground objects from the background, as well as distinguish between 'old' and 'recent' background. That is, an object which has reached a stable position in the scene (a recently parked car, a newly abandoned bag, a moving person becoming still), which tends to become part of the background in most state of the art techniques, can be easily identified with the proposed approach.

The future work will be in the direction of extending the region-based decisions to other parts of the system (e.g. shadow detection, currently a pixel based technique is used [5]) and developing scene understanding techniques based on the presented segmentation and tracking system.

## ACKNOWLEDGEMENTS

This work has been supported by Spanish Project TIC2001-0996 and by the IST programme of the EU IST-2000-32795 SCHEMA.

## REFERENCES

[1] S. Jabri, Z. Duric, H. Wechsler & A. Rosenfeld, "Detection and location of people in video images using adaptive fusion of color and edge information," *Proc.of ICPR'2000*.

[2] C. Stauffer, W.E.L. Grimson, "Learning patterns of activity using real-time tracking," *IEEE trans. on Pattern Analysis and Machine Intelligence*, **22**(8), August 2000.

[3] A. Elgamal, R. Duraiswami, D. Harwood and L. Davis, "Background and foreground modeling using nonparametric Kernel density estimation for visual surveillance," *Proceedings of the IEEE*, **90**(7), July 2002.

[4] Q. Zhou and J.K. Aggraval, "Tracking and classifying moving objects from video," *Proc. of PETS'2001*, Hawaii, 2001.

[5] J.L. Landabaso, L.Q. Xu, M. Pardàs, "Robust tracking and object classification towards automated video surveillance", *Proceedings of Internacional Conference on Image Analysis and Recognition* (ICIAR 2004), Porto, Portugal, September 2004.

[6] P. Salembier, L. Garrido "Binary partition tree as an efficient representation for image processing and information retrieval", *IEEE Trans. on Image Processing*, **9**(4):561-576, Apr. 2000