Region-Oriented Visual Attention Framework for Activity Detection

Thomas Geerinck and Hichem Sahli

Electronics & Informatics Department - VUB-ETRO
Vrije Universiteit Brussel (VUB)
Interdisciplinary Institute for BroadBand Technology (IBBT),
Brussels, Belgium
http://www.etro.vub.ac.be

Abstract. This paper proposes a framework, based on a spatio-temporal attentive mechanism, for automatic region-of-interest determination, corresponding to events in video sequences of natural scenes of dynamic environments. We view this work as a preliminary step towards the solution of high-level semantic event analysis. More specifically, we wish to detect a visual event within a cluttered scene, without intensive training algorithms. In contrast to event detection methods used in the literature, which drive attention based on motion and spatial location hypothesis, in our approach the visual attention is region-driven as well as feature-driven. For this purpose, a two stages attention mechanism is proposed. In a first phase, spatio-temporal activity analysis extracts key-frames from the image sequence and selects salient areas within these frames. The three types of visual attention features are used, namely, intensity, color and motion. Consequently, the selected areas are further processed to determine the most active region, based on a newly defined region saliency measure. Qualitative and quantitative results, using the proposed framework, are illustrated envisaging the application domain of change detection in automated visual surveillance.

Keywords: Event detection, activity measure, visual attention, region-oriented.

1 Introduction

It is common that, we as humans shift our attention toward anything that is interesting to us. This process is a natural form of perception of event/activity detection. In general terms, an event can be defined as a qualitatively significant change in the behavior of the data, as defined by the domain user [1]. In the context of video analysis, a visual event is commonly defined based on a moving object with constraints in its size, color, shape, or motion instances that haven't been seen before [2] [3].

Our goal in this work is to formalize a strategy for efficient detection and localization of target region activity or event (region-of-interest), from video sequences of natural scenes of dynamic environments. This issue forms the basis

L. Paletta and E. Rome (Eds.): WAPCV 2007, LNAI 4840, pp. 481–496, 2007.

of what might lead in a later stage to object behavior recognition and understanding [4], which is yet an unsolved problem.

In this work, biologically inspired methods have been chosen. Attentive mechanisms are found to be relevant, e.g., for the selection of incoming visual information, for the decision making in top-down, i.e., symbolic to sensory information processing, for the selective functioning within the organization of behaviors, or for the understanding of individual and social cognition [5]. Attentive mechanism, in computer vision, aims at mimicking the ability of natural vision systems to select just the relevant aspects of the broad visual input, and should be considered as a set of strategies that attempts to reduce the computational cost of the search processes inherent in visual perception [6]. For modeling visual attention, it is crucial to select an appropriate set of visual descriptors, e.g. local color descriptors, color histograms, and motion descriptors, which can help establishing a connection towards the semantically meaningful features of content [7] [8].

Conventional region-of-interest (ROI) determination based on visual attention principles, encapsulates (i) temporal and motion information which characterize the selected event, (ii) static or video-based feature combination method, and (iii) integrating saliency-oriented and task-oriented [9]. Motion is indeed of fundamental importance in biological vision systems and contributes to visual attention as confirmed by Watanabe et al. [10]. As such motion will be the most important cue which will be used in the proposed framework. Note that, our current implementation, considers only saliency-oriented ROI determination, meaning no top-down influences are reflected.

In our approach, events are regarded as qualitatively significant changes in the behavior of a defined motion activity measure [11,8]. In contrast of previously defined motion activity measures, e.g. as in [8], where two simple descriptors have been used for describing monotonous activity (defined as the average block-based motion vector magnitude) and non-monotonous activity (approximation as the average temporal derivative of motion vectors), in this work we combine intensity, color, and motion features. This allows defining motion activity as the gross, overall motion content in a given video segment, such as, high or low activity, spatially coherent or scattered activity, etc... In contrast to event/novelty detection methods used in the literature, e.g. [2] where a clustering based learning mechanism that incorporates habituation theory is applied, our approach detects valuable events in cluttered and chaotic environments without intensive training algorithms. These characteristics make our approach highly valuable envisaging change detection in automated visual surveillance as application domain.

Figure 1 depicts the proposed framework for region-based event detection. It is composed of two major modules:

- Activity detection module (Section 2.1), consisting of a spatio-temporal motion processing module allowing the detection of key-frames in the image sequence, and
- A region-driven focus of attention module (Section 2.2). Attention allows us to focus on the relevant regions in the scene and thus reducing the amount of information needed for further processing.

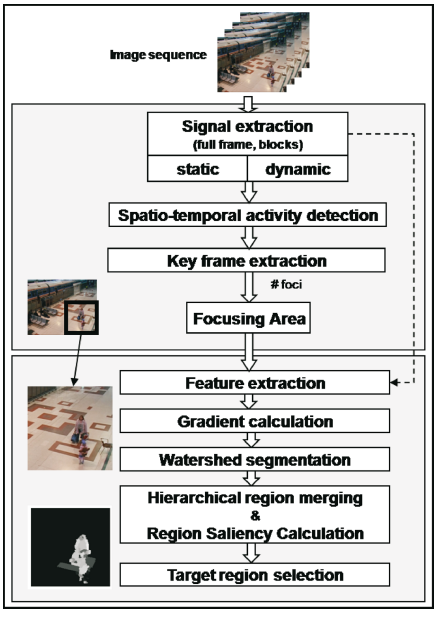


Fig. 1. Global system architecture

Qualitative and quantitative results are given in Section 3. Finally, Section 4 draws some conclusions and outlines further research.

2 Event Detection in Video Streams

Consider a dynamic phenomenon, whose behavior changes enough over time so as to be considered a qualitatively significant change. Each such change is an event. An example is the change of Station Hall traffic from normal to someone running. In order to detect an event, outliers, peaks, trends or changes in trends should be spotted automatically in the available measures (data) considering a well specified time-window. Using a short time-window, instantaneous events or activities are emphasized, in general represented by outliers or sudden change in amplitude of the signal. Consequently, using a longer time-window, time series analysis techniques can be applied to detect other types of peculiarities, such as trends, changes in trends, periodicity, etc.

In this work, we focus on detecting frame-to-frame events, possible starting points of an event, etc... Events lasting for a certain time period or during a video segment, are not considered at this stage of the development. In this section, we first define the *activity measure* used as time series data for event analysis, and second we propose methods for event detection by identifying the time points at which changes, in the activity measure, occur.

2.1 Activity Measure

In the current framework only short time-windows are considered. We define a time-window T_w of approximately 3 - 4 s. Given an RGB image sequence with a frame rate f of 20 - 30 Hz, the analysis is made for $N = fT_w$ frames. Each image frame, $F_k(k = 1 \dots N)$, is divided into n quadrants, $F_k^j(j = 1 \dots n)$ (100 in the current implementation). For each quadrant j an activity measure over time, $A_k^j; k = 1 \dots N$, is estimated. Time series analysis will allow detecting the key-frame containing an event, and to localize the detected event within the frame, for further analysis.

The activity measure, within an RGB image quadrant F_k^j, includes the estimation of:

- Δh_k^j the inter-frame color histogram change,
- $\mu(\Delta\mathcal{I})_k^j$ the inter-frame mean intensity change,
- $\mu(\mathcal{M})_k^j$ the mean motion vector,
- $\mu(\dot{\mathcal{M}})_k^j$ the mean acceleration, estimated as mean difference in motion.

The HSI color space has been chosen to provide an intuitive representation of color and to approximate the way in which humans perceive and manipulate color [13]. Let $p = [x, y]^t$ the pixel coordinates. We define $Max = max(\mathcal{R}(p), \mathcal{G}(p), \mathcal{B}(p))$, and $Min = min(\mathcal{R}(p), \mathcal{G}(p), \mathcal{B}(p))$. The RGB to HSI transformation is defined as follows:

$$\mathcal{H}(p) = \begin{cases} undefined & , \quad if Max = Min \\ 60\frac{\mathcal{G}(p)-\mathcal{B}(p)}{Max-Min} & , if Max = \mathcal{R}(p) \bigwedge \mathcal{G}(p) \geq \mathcal{B}(p) \\ 60\frac{\mathcal{G}(p)-\mathcal{B}(p)}{Max-Min} + 360 & , if Max = \mathcal{R}(p) \bigwedge \mathcal{G}(p) < \mathcal{B}(p) \\ 60\frac{\mathcal{G}(p)-\mathcal{B}(p)}{Max-Min} + 120 & , \quad if Max = \mathcal{G}(p) \\ 60\frac{\mathcal{G}(p)-\mathcal{B}(p)}{Max-Min} + 240 & , \quad if Max = \mathcal{B}(p) \end{cases}$$

$$\mathcal{S}(p) = \begin{cases} 0 & , if Max = 0 \\ 1 - \frac{Min}{Max} & , \ otherwise \end{cases}$$

$$\mathcal{I}(p) = Max$$

For the color histogram estimation, the HSI space is uniformly quantized into a total of 256 bins. This includes 16 levels for $\mathcal{H}$, 4 levels for $\mathcal{S}$, and 4 levels for $\mathcal{I}$. Finally, for an image quadrant j, the inter-frame color histogram distance, Δh_k^j , is estimated using the Euclidean distance between the color histograms at time k and $k + 1$, defining a color similarity [14] between successive image frames.

The inter frame change $(\Delta\mathcal{I})_k$ is estimated as $|\mathcal{I}_k - \mathcal{I}_{k+1}|$; $\mathcal{I}_k$ being the intensity map.

The motion map $\mathcal{M}$ correspond to the modulus of the *optical flow*, estimated using the Lucas-Kanade [12] algorithm applied on the intensity map $\mathcal{I}_k$. The mean acceleration of quadrant j at time k, $\mu(\dot{\mathcal{M}})_k^j$, is estimated as the mean (over all the pixels in quadrant j) difference between successive motion maps $\mathcal{M}_k$ and $\mathcal{M}_{k+1}$.

For each quadrant j, the activity measure at time k, is given by:

$$A_k^j = \mu(\mathcal{M})_k^j[1 + Peak(\mu(\Delta\mathcal{I})_k^j) + \mu(\dot{\mathcal{M}})_k^j] + Peak(\Delta h_k^j) \tag{1}$$

with

$$Peak(X) = \begin{cases} X \ if|X - \mu(X)| \geq \alpha\sigma(X) \\ 0 \ \ if|X - \mu(X)| < \alpha\sigma(X) \end{cases} \quad (2)$$

where $X = \{\mu(\Delta\mathcal{I})_k^j, \Delta h_k^j\}$, $\mu(X), \sigma(X)$ are the mean and standard deviation, respectively, estimated for $k = 1 \ldots N$, and $\alpha = 0.5$ (in our experiments).

As it will be seen in Section 3, the behavior of the time series data, given by the proposed activity measure, fits our objective for detecting frame-to-frame events or possible starting points of an event. These events are represented by a peak (maxima) in the activity, or a sudden activity change. In other words, when the motion in a particular image quadrant stays monotonous over a certain time window, the measured activity will be low, or even 0, due to the influence of the $Peak()$ function. On the other hand if an object is rapidly passing through a certain quadrant, a maximum in the activity signal of that particular quadrant is generated.

2.2 Event Analysis

The specific problem we address here is the identification of the time points at which changes occurs in the time series A_k^j. These change-points are referred to as frames of interest or key-frames. Note that, we consider not only the detection of frames of interest, but also the most active area (focusing area) within the frames of interest. Two approaches are proposed:

- Peak Detection, for which we consider the competition between the activity measures from all image quadrants and select the highest peak. An inhibition mechanism, on the activity measures, is incorporated in order to avoid reselection
- Change-Point Detection, for which we consider a frame activity measure (by combining the quadrant's activity measure) from which change-points are estimated by finding the best set of points that minimizes the error in fitting a pre-defined function. The appropriate set of points is found based on maximum likelihood method [16]. From the detected key-frames, the quadrant which has the highest activity value is selected

2.2.1 Peak Detection

Given the above activity measure for all the image quadrants and for all the frames in the defined time window, competition for focus of attention is conducted by selecting the highest peaks of activity in the spatial as well as temporal domain. The selection of the winning peak results in selecting an image quadrant in a given frame of the time window. The actual focusing area, is constituted by the selected quadrant and its surrounding quadrants. If again in one of the surrounding quadrants the activity measure is high enough, the corresponding surrounding quadrants are added to the focusing part of the frame, as illustrated in Figure 2.

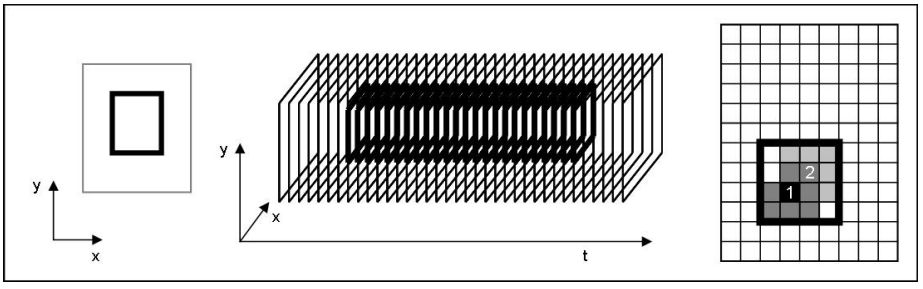


Fig. 2. Principle of inhibition of return in spatial and temporal domain to avoid reselection

Comparable to what has been proposed in existing visual attention simulating architectures [5] [15], a principle of inhibition of return is integrated, in the temporal as well as the spatial domain to avoid reselection of the same area. The result of this approach is illustrated in Figure 2, where the focused part of the image is shown as a bold rectangle. Consequently, the attended area of interest is inhibited in the previous l frames as well as the following l frames, where l depends on the length of the time-window (T_w) and on the frame rate. More concretely, inhibition in our approach is defined as a suppression of the activity measure in the focused area, for a specified amount of time.

2.2.2 Change-Point Detection

In this approach we follow the work of Guralnik and Svrivastava [16], where statistical change-point detection for time series analysis has been proposed. In standard statistical approaches, change-point detection has been made by (a) *apriori* determining the number of change-points to be discovered, and (b) deciding for the model to be used for fitting the subsequence between successive change-points. In [16] a method has been proposed for the detection of the appropriate set of number of points that minimizes the error in fitting a predecided function using maximum likelihood. Moreover, in [16], no constraints are imposed on the class of functions that will be fitted to the subsequences between successive change-points. Two approaches have been proposed, the batch (offline) and the incremental (online). In the batch version, the entire data set is available, from which the best set of change-points is determined. In the incremental version, the algorithm receives new data points one at a time, and determines if the new observation causes a new change-point to be discovered.

Following the notation in [16], let $A_k(k = 1, ..., N)$ be the time series to be segmented. Here k is the time variable, and A_k corresponds to the frame activity estimated as a weighted sum of the quadrants activities $A_k^j(j = 1, ..., n)$. The weight assigned to each quadrant is the standard deviation of the quadrant's activity.

The change-points detection, is then formulated as finding a piecewise segmented model, given by

$$\begin{aligned} A &= f_1(k, \mathbf{w_1}) + e_1(k), (1 < k \leq \theta_1), \\ &= f_2(k, \mathbf{w_2}) + e_2(k), (\theta_1 < k \leq \theta_2), \\ &= ..... \\ &= f_l(k, \mathbf{w_l}) + e_l(t), (\theta_{l-1} < k \leq N). \end{aligned}$$

Where $f_i(k, \mathbf{w_i})$ is the function (with its vector of parameters $\mathbf{w_i}$) that is fitted to the segment i. The θ_i's are the change-points between successive segments, and $e_i(t)$'s are error terms. Several type of basis functions can be considered, e.g. algebraic polynomials, wavelet, Fourier, etc...[17]. In our implementation, algebraic polynomials of the form given by Eq. 3 are considered.

$$p(x) = p_1 x^m + p_2 x^{m-1} + \ldots + p_m x + p_{m+1} \tag{3}$$

The reader is referred to [16] for the detailed maximum likelihood estimation method for the detection of the change-points θ_i. In our case, every change-point corresponds to a frame of interest or key-frame. Within each key-frame, the area with highest activity value is selected as first focusing area. The parameters vector $\mathbf{w_i}$ consist of :

- m: the polynomial degree
- p: the minimal number of points, in each segment, required for the model fitting

A third parameter affecting the behavior of the algorithm is

- δ: a user-defined threshold defining a stopping criteria in the case of the batch version, and a likelihood increase threshold for the incremental version, respectively.

2.3 Region-Based Visual Attention

Having detected the key-image frame with high/novel activity, as well the image quadrants where the activity has been detected, we can now determine in more focused way the meaningful region of interest. The objective is to partition the selected quadrants into disjoint regions, in a manner consistent with human perception of the content. We propose a region-based attentional selectivity, with the following steps: (1) feature extraction, (2) initial region-based segmentation, (3) region merging, (4) region saliency calculation, and (5) target meaningful-region selection.

2.3.1 Feature Extraction

From the RGB color image quadrant inputs a set of feature maps are estimated (a) two color contrast $\mathcal{I}_{\mathcal{RG}} = |\mathcal{R} - \mathcal{G}|$ and $\mathcal{I}_{\mathcal{BY}} = |\mathcal{B} - \mathcal{Y}|$ (with $\mathcal{Y} = (\mathcal{R} + \mathcal{G})/2 - (\mathcal{R} - \mathcal{G})/2 - \mathcal{B}$), (b) an intensity contrast $\mathcal{I}$, and (c) a motion map $\mathcal{M}$.

2.3.2 Initial Region-Based Segmentation

For the region-based segmentation the watershed method is used. The watershed transform [18] is a morphological segmentation tool often applied on the gradient magnitude of an image in order to guide the watershed lines to follow the crest lines and the real boundaries of the objects. In grey level images, the modulus of the gradient is a scalar function of the coordinates and expresses the distance between neighboring pixels with respect to their intensity. In our case, considering the 4 feature maps, we obtain a spectral image, for which the modulus of the gradient expresses the distance in the chosen spectral space using a certain metric. Here, we follow the approach proposed in [20]. The spectral gradient is estimated as a weighted sum of the individual gradients:

$$GS(x,y) = \varepsilon_1|\nabla\mathcal{I}_{\mathcal{RG}}(x,y)| + \varepsilon_2|\nabla\mathcal{I}_{\mathcal{BY}}(x,y)| + \varepsilon_3|\nabla\mathcal{I}(x,y)| + \varepsilon_4|\nabla\mathcal{M}(x,y)| \quad (4)$$

The output of the watershed segmentation is a Region Adjacency Graph, $G(P^0, E)$. The nodes, $P^0 = \{r_1^0, r_2^0, ..., r_{m_0}^0\}$ are the set of regions, and the arcs E connecting the nodes are the the boundaries between neighboring regions.

The advantageous characteristics of applying the watershed transform are : (i) The fact that watersheds form closed curves, providing a full partitioning of the image domain, thus it is a pure region-based segmentation which does not require any closing or connection of the edges, (ii) Watersheds of the gradient magnitude can play the role of a multiple point detector, thus treating any case of multiple region coincidence [19], (iii) there is a 1-1 relationship between the minima and the catchment basins. This latter characteristic is used in the region merging process of the next section.

2.3.3 Region Merging

A meaningful image segmentation groups the pixels into disjoint regions that consist of uniform components. Facing absence of contextual knowledge, the only alternative which can enrich our knowledge concerning the significance of our segmented groups is the creation of a hierarchy guided by the knowledge which emerges from the superficial and deep image structure. Our main goal here, is to create a hierarchy among the gradient watersheds which preserves the topology of the initial watershed lines and extracts homogeneous objects of a larger scale.

Let us first define what we mean by a hierarchy. Let $P^0 = \{r_1^0, r_2^0, ..., r_{m_0}^0\}$ be the initial partitioning of the image F_k obtained by applying the watershed transformation on the gradient image (Eq.4). A *hierarchical level* h (HL_h) is defined as the partitioning $P^h = \{r_1^h, r_2^h, ..., r_{m_h}^h\}$ which preserves the inclusion relationship $P^h \supseteq P^{h-1}$, implying that each atom of the set P^h is a disjoint union of atoms from the set P^{h-1}.

In our approach, we create such a hierarchy using the *waterfall* algorithm proposed in [21]. In this algorithm, a hierarchy is constructed by successively merging neighboring regions according to a saliency measure associated to the common contour (arc), estimated as the relative altitude of the regional minima. In our implementation, the saliency measure of an arc is adapted throughout the

hierarchy. At hierarchical level h, a saliency, $c(e_{ij}^h)$, is associated to the arc e_{ij}^h connecting two neighboring regions r_i^h and r_j^h:

$$c(e_{ij}^h) = g_{e_{ij}^0} + s(r_i^h, r_j^h) \tag{5}$$

where

- $g_{e_{ij}^0}$ is the height of the saddle point of the contour, being the lowest gradient value $GS(x, y)$ of the pixels forming the boundary between the neighboring regions r_i^h and r_j^h.
- $s(r_i^h, r_j^h) = max(|S(r_i^h) - S(r_i^h \bigcup r_j^h)|, |S(r_j^h) - S(r_i^h \bigcup r_j^h)|)$; with $S(r) = \mu_{\mathcal{I}}(r) + \mu_{\mathcal{I}_{\mathcal{RG}}}(r) + \mu_{\mathcal{I}_{\mathcal{BY}}}(r)$. $\mu_{\mathcal{I}}, \mu_{\mathcal{RG}}$, and $\mu_{\mathcal{BY}}$, being the mean intensity contrast, and the mean color contrasts.

In our current implementation, the hierarchical segmentation, and thus the formation of new levels, stops when the number of regions is reduced to 25%. Note that, during the hierarchical levels retrieval, the Region Adjacency Graph is updated.

2.3.4 Region Saliency Value

Let h be the retrieved hierarchical level. The goal of this step is to associate to each region r_i^h, of the partition P^h, a single-valued saliency, representing the level of interest or the significance of the region with respect to its surrounding regions. The region saliency is defined as:

$$Sal(r_i^h) = Card(r_i^h) d(r_i^h; 0) \mathcal{P}(r_i^h) \mu_{\mathcal{M}}(r_i^h) Csr(r_i^h) \tag{6}$$

with

- $Card(r_i^h) = |r_i^h|$ is the number of pixels of the region,
- $d(r_i^h; 0)$ is the Euclidian distance between the center of gravity of the region and the image center,
- $\mathcal{P}(r_i^h)$ is the region homogeneity [22]. It is defined as a composition of two components, standard deviation and discontinuity of intensities. Standard deviation describes the contrast within a local region. Discontinuity, is a measure of abrupt changes in gray levels obtained using the gradient of the image. Thus, $\mathcal{P}(r) = 1 - \sigma(r)V(r)$ with $\sigma(r) = 1/|r|(\sum_{(x,y)\in r}[\mathcal{I}(x, y) - \mu_{\mathcal{I}}(r)]^2)^{1/2}$ and $V(r) = \sum_{(x,y)\in r} |\nabla\mathcal{I}(x, y)|$. $\mu_{\mathcal{I}}(r)$ being the mean intensity value of the region r. Note that, σ and V are normalized in order to have the value of the homogeneity in the range $[0, 1]$.
- $\mu_{\mathcal{M}}(r_i^h)$ is the mean motion of the region r_i^h,
- $Csr(r_i^h)$ is the *Center-Surround* operation between the focused region r_i^h and the neighboring regions. We have extended the well known center-surround operation on pixels [5] towards regions. $Csr(r)$ should indicate the importance or the level of significance of r with respect to its adjacent regions $r_1, \ldots r_l$. The idea here is to compare the importance of r and the hypothetical merged region $R = r \bigcup r_1 \ldots \bigcup r_l$. Thus:

$$Csr(r) = |\mu_{\mathcal{I}}(r) - \mu_{\mathcal{I}}(R)| + |\mu_{\mathcal{RG}}(r) - \mu_{\mathcal{RG}}(R)| + |\mu_{\mathcal{BY}}(r) - \mu_{\mathcal{BY}}(R)| + |\mu_{\mathcal{H}}(r) - \mu_{\mathcal{H}}(R)|$$

$\mu_{\mathcal{I}}, \mu_{\mathcal{RG}}, \mu_{\mathcal{BY}}, \mu_{\mathcal{H}}$ being the mean intensity $\mathcal{I}$, the mean color contrast $\mathcal{I}_{\mathcal{RG}}$, $\mathcal{I}_{\mathcal{BY}}$, and the mean hue ($\mathcal{H}$), respectively.

Note that, all the factors in Eq. 6 are normalized. The latter equation is used for the estimation of the active-region. A scan-path of target regions, as described in [23], can be determined consequently since all information is available.

3 Experimental Results

The proposed framework, for automatic active-region determination, has been tested on two image sequences: (i) the Munich train station (provided by the Institute of microtechnology, University of *Neuchâtel*, Switzerland), and (ii) the PETS 2007 (scene 4, camera 2) sequence [24].

In order to assess our results, "ground truth" events have been defined by a specialized surveillance company, highlighted with ellipses in Figure 3 and Figure 4. The following events have been defined. For the train station sequence (Figure 3), the person dressed in red, running in the middle of the hall in frame 50; the two tall persons appearing in the scene in the left bottom corner in frame 60 and 80; the person walking in the middle of the hall in frame 90; the two persons reappearing from behind a pillar in the middle of the scene in frame 30; the person dressed in black, reappearing from behind the publicity panel near the stairs in frame 90.

For the PETS 2007 sequence (Figure 4), the events include, the group of people entering the hall in the right top corner in frame 10; the person together with 2 children, who suddenly start running all in frame 50; the tall guy appearing from the left bottom corner and going towards the center of the camera view in frame 80; the person dressed in red appearing and disappearing in the right top corner in frame 60.

In the following we first discuss the proposed framework using the train station sequence, and then we illustrate the obtained results of the PETS 2007 sequence. For both sequences, a time-window (N) of 100 frames has been considered.

3.1 Event Detection

Figure 5 depicts the activity measure (Eq 1) over time for several image quadrants, as well as the winning quadrants (highest peak in the curve) using the proposed Peak Detection Method and its associated inhibition of return. Iteratively, the highest peak (considering all the quadrants) is selected, and consequently its corresponding activity measure is inhibited, meaning reduced considerably in amplitude. The detected frames of interest, as well as the selected (winning) active-quadrant are shown in Figure 6.

For the Change-Point Detection experiments, a polynomial (Eq. 3) of degree $m = 5$ has been used, and the number of points needed for the model fitting in

Fig. 3. The test sequence of Munich train station

Fig. 4. The PETS 2007 test sequence

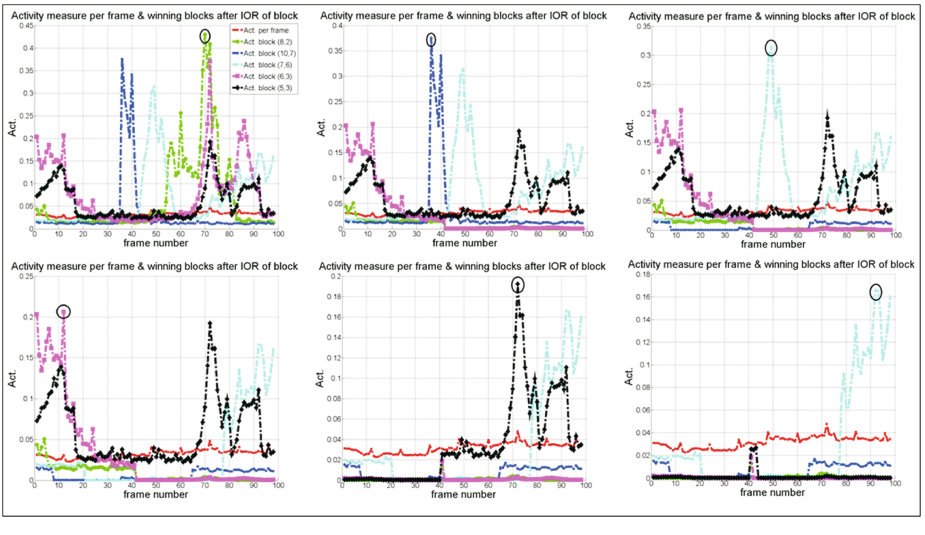


Fig. 5. The quadrants activity measure and winning quadrants

Fig. 6. Peak Detection Method Results

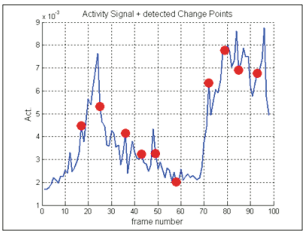
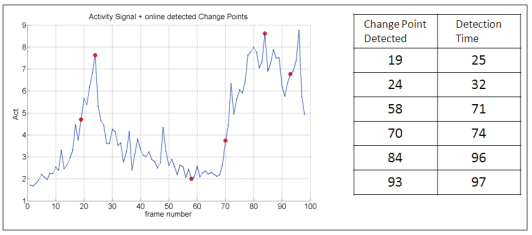


Fig. 7. Change-Point Detection - Batch version: Activity measure and detected change-points

each segment has been set to $p = 5$. These values have been chosen empirically. Figure 7 and Figure 8 show the obtained results using the batch version, and Figure 9, and Figure 10 depict the results obtained using the incremental version.

Fig. 8. Change-Point Detection - Batch version: Key-frames and associated active-quadrant

Comparing the results of Figure 10 and the ground truth of Figure 3 one can notice that the proposed activity measure and the proposed change-point detection algorithm allow detecting the main events in the image sequence.



Change Point Detected	Detection Time
19	25
24	32
58	71
70	74
84	96
93	97

Fig. 9. Change-Point Detection - Online version: Activity measure and change-points

Fig. 10. Change-Point Detection - Online version: Key-frames and associated active-quadrant

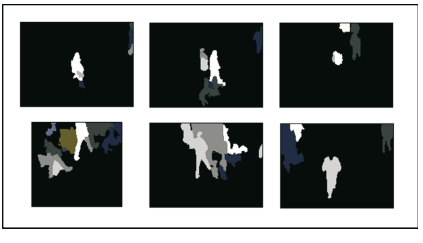


Fig. 11. The extracted active-regions in the detected key-frames

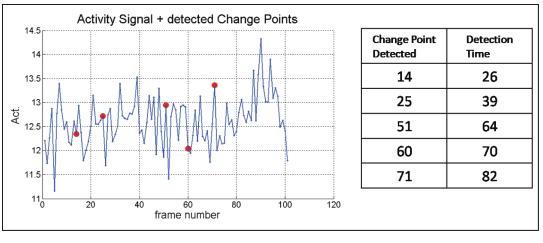
3.2 Active-Region Determination

The proposed region-based visual attention mechanism of Section 2.3 has been applied to the key-frames of Figure 10. Figure 11 depicts the region saliency map as defined in Eq. 6, where bright values correspond to high salient regions. One can notice that the segmentation produces meaningful regions, following the shape of the perceptual object. Moreover, comparing the detected image frames (quadrants) and the salient regions to the visually selected events (Figure 3), one can notice that the proposed activity detection module with online event

detection method selected (a) the running person with red jacket, (b) the two persons appearing in the scene at the left bottom, (c) the person walking in the middle of the hall, as well as (d) the two persons reappearing from behind a pillar in the middle of the scene.

3.3 Additional Results

Concerning the Change-Point Detection experiments on the PETS 2007 image sequence, a polynomial (Eq. 3) of degree $m = 4$ has been used, and the number of points needed for the model fitting in each segment has been set to $p = 5$. These values have been chosen empirically. Figure 12, and Figure 13 depict the results obtained using the incremental version. The proposed region-based visual attention mechanism of Section 2.3 has been applied to the active-quadrants. Figure 13 depicts the region saliency map associated to the selected active-quadrants, as defined in Eq. 6, where bright values correspond to high salient regions. Again, one can notice that the segmentation produces meaningful regions, following the shape of the perceptual object. As such, our proposed method selected as active regions (a) several persons entering the hall at the top right corner, (b) the person dressed in red, (c) the running person, with two children, (d) the tall person moving towards the center of the camera view.



Change Point Detected	Detection Time
14	26
25	39
51	64
60	70
71	82

Fig. 12. Change-Point Detection - Online version: Activity measure and change-points

Fig. 13. Change-Point Detection - Online version: Key-frames, associated active-quadrant and extracted active-regions

4 Conclusions

In this paper we addressed the problem of activity detection in cluttered scenes, as posed in the application domain of automated visual surveillance. We proposed a new region-based spatio-temporal attentive mechanism. The main contributions of this work include

- The definition of a spatio-temporal activity measure
- The use of general approaches to change-point detection, i.e. event detection, that do not require training
- The development of a region-based focus of attention mechanism integrating spatio-temporal features

Comparing the results of the proposed framework with visual event (change-point) detection by humans gives promising results.
The principal limitations of our approach exist mainly in the low-level segmentation used to drive the determination of an active region and in a later stage, the constitution of a meaningful object. Therefore, future research will include (i) combining arc saliency and region saliency for a better region segmentation, (ii) develop an objective segmentation evaluation method defining a stopping criteria for the hierarchical segmentation taking into account the image content, and (iii) include a high-level semantic event analysis to classify the detected event.

Acknowledgments. This work has been partly funded by (i) the Institute for the Promotion of Innovation by Science and Technology in Flanders(IWT) and BARCO-Belgium under in the framework of the ITEA project SERKET - Securisation Keeps Threats, and (ii) the IBBT-VIN project, co-funded by IBBT.

The authors would like to thank Dr. N. Ouerhani, from the Institute of microtechnology, University of *Neuchâtel*-Switzerland, for the provided video sequence and the valuable discussion.

References

1. Varadharajan, C.: A Wavelet-Based System for Event Detection in Online Real-time Sensor Data. Massachusetts Institute of Technology (2004)
2. Gaborski, R.S., Vaingankar, V.S., Chaoji, V.S., Teredesai, A.M.: A System for Novelty Detection in Video Streams with Learning. Laboratory for Applied Computing, Rochester Institute of Technology, Rochester, NY, USA (2004)
3. Tentler, A., Vaingakar, V.S., Gaborski, R.S., Teredesai, A.M.: Event detection in video sequences of natural scenes. Rochester Institute of Technology, Laboratory for Applied Computing (2002)
4. Tsotsos, J.K.: Distributed Saliency Computations Solve the Feature Binding Problem. In: Proc. ECCV WAPCV, Prague (May 15, 2004)
5. Itti, L.: Models of Bottom-Up and Top-Down Visual Attention. Ph.D. Thesis, California Institute of Technology (2000)
6. Tsotsos, J.K.: Motion Understanding: Task-Directed Attention and Representations that link Perception with Action. International Journal of Computer Vision 45(3), 265–280 (2001)

7. Rapantzikos, K., Avrithis, Y., Kollias, S.: On the use of spatiotemporal visual attention for video classification. In: VLBV 2001. Proc. of Int. Workshop on Very Low Bitrate Video Coding (2005)
8. Peker, K.A., Alatan, A.A., Akansu, A.N.: Low-level motion activity features for semantic characterization of video. ICME 2000 2, 801–804 (2000)
9. Hu, Y., Xie, X., Ma, W-Y., Chia, L-T., Rajan, D.: Salient region detection using weighted feature maps based on the human visual attention model. In: Aizawa, K., Nakamura, Y., Satoh, S. (eds.) PCM 2004. LNCS, vol. 3331, Springer, Heidelberg (2004)
10. Rapantzikos, K., Tsapatsoulis, N.: Enhancing the robustness of skin-based face detection schemes through a visual attention architecture. ICIP 2005 II, 1298–1301 (2005)
11. Makrogiannis, S.K., Bourbakis, N.G.: Motion analysis with application to assistive vision technology. In: Makrogiannis, S.K., Bourbakis, N.G. (eds.) ICTAI 2004. 16th IEEE International Conference on Tools with Artificial Intelligence, pp. 344–352 (2004)
12. Lucas, B.D., Kanade, T.: An iterative image registration technique with an application to stereo vision. In: Lucas, B.D., Kanade, T. (eds.) International Joint Conference on Artificial Intelligence, pp. 674–679 (1981)
13. Manjunath, B.S., Ohm, J-R., Vasudevan, V.V., Yamada, A.: Color and texture descriptors. IEEE Transactions On Circuits And Systems For Video Technology 11(6), 703–715 (2001)
14. Smith, J.R., Chang, S-F.: Tools and Techniques for Color Image Retrieval. Storage and Retrieval for Image and Video Databases (SPIE) , 426–437 (1996)
15. Tsotsos, J.K., Culhane, S., Wai, W., Lai, Y., Davis, N., Nuflo, F.: Modeling visual attention via selective tuning. Artifical Intelligence 78(1-2), 507–547 (1995)
16. Guralnik, V., Srivastava, J.: Event detection from time series data. In: Guralnik, V., Srivastava, J. (eds.) KDD 1999. Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining, San Diego, California, United States, pp. 33–42 (1999)
17. Cherkassky, V., Mulier, F.: Learning from Data. Wiley-Interscience, New York, NY, USA (1998)
18. Meyer, F.: An Overview of Morphological Segmentation. IJPRAI 15(7), 1089–1118 (2001)
19. Vanhamel, I., Pratikakis, I., Sahli, H.: Multiscale gradient watersheds of color images. IEEE Transactions on Image Processing 12(6), 617–626 (2003)
20. O'Callaghan, R.J., Bull, D.R.: Combined morphological-spectral unsupervised image segmentation. IP 14(1), 49–62 (2005)
21. Marcotegui, B., Beucher, S.: Fast implementation of waterfall based on graphs. In: Ronse, C., Najman, L., Decenciere, E. (eds.) Mathematical morphology: 40 years on. Proceedings of the 7th international symposium on mathematical morphology. Computational imaging and vision, vol. 30, pp. 177–186 (2005)
22. Cheng, H-D., Sun, Y.: A Hierarchical approach to color image segmentation using homogeneity. IEEE Transactions on Image Processing 9(12), 2071–2082 (2000)
23. Sun, Y., Fisher, R.: Object-based Visual Attention for Computer Vision. Artificial Intelligence , 77–123 (2003)
24. `http://ftp.pets.rdg.ac.uk/`