A Bottom-Up Spatiotemporal Visual Attention Model for Video Analysis

Konstantinos Rapantzikos[1], Nicolas Tsapatsoulis[2], Yannis Avrithis[1],

Stefanos Kollias[1]

[1] **School of Electrical & Computer Engineering, National Technical University of Athens**

[2] **Department of Computer Science, University of Cyprus**

Correspondences to:

Konstantinos Rapantzikos

School of Electrical & Computer Engineering

National Technical University of Athens

Hroon Polytexneiou 9, 15773

Tel: +302107724351

Fax: +32107722492

Email: **rap@image.ntua.gr**

1. ABSTRACT

A video analysis framework based on spatiotemporal saliency calculation is presented. We propose a novel scheme for generating saliency in video sequences by taking into account both the spatial extent and dynamic evolution of regions. Towards this goal we extend a common image-oriented computational model of saliency-based visual attention to handle spatiotemporal analysis of video in a volumetric framework. The main claim is that attention acts as an efficient preprocessing step of a video sequence in order to obtain a compact representation of its content in the form of salient events/objects. The model has been implemented and qualitative as well as quantitative examples illustrating its performance are shown.

1. INTRODUCTION

Primate vision provides natural solutions to many machine vision problems. If it were possible to embody them in a computational theory, then machine vision would be successful. Recently, a central part of the human vision system (HVS), namely the ability to concentrate of salient regions of the visual input, has attracted several researchers both from the field of neuroscience and computer vision. This ability of the HVS states that despite the common belief that we see everything around us, only a small fraction of the surrounding visual information is processed at any time and leads to higher level understanding of the visual scene. One of the dominant theories in the field is *saliency-based visual attention* (VA) [17, 21].

Complete vision application systems invoke attentional mechanisms in order to confront the computational load of several higher level processing steps [8, 35]. If the attended regions represent the input well, a great deal of search can be avoided. Two

major attentional mechanisms are known to control the visual selection process. First, bottom-up attentional selection is a fast, and often compulsory, stimulus-driven mechanism. Involuntary attention capture by distracting inputs occurs only if they have a property that a person is using to find a target [31]. Second, top-down attentional selection initiates from the higher cognitive levels in the brain that influence the attentional system to bias the selection in favor of a particular (or a combination) of feature(s). Only information about the region that is preattentively extracted can be used to change the preferences of the attentional system.

In the field of computational video analysis, image sequences are usually processed and analyzed in a frame-by-frame basis in order to infer the short-term objects' temporal evolution. Such methods use information over a small number (typically two) of frames. Linking together the obtained results generates longer-term dynamics. The actual long-term temporal dimension of the video data is therefore disregarded by incorporating parametric motion model assumptions or smoothing constraints. Such methods are prone to noise and can lead to high computational complexity if e.g. accurate motion estimation is one of the prerequisites [2, 16]. For video analysis, it is beneficial to use spatiotemporal filtering of a larger neighborhood (a volume of data) in order to include the informative temporal dimension. Adelson and Bergen [1] were the first to suggest computational solutions for spatiotemporal filtering of volumetric data. Bolles and Baker [4, 3] also exploit spatiotemporal volumes to extract motion parameters of a camera, moving in a straight line, using epipolar plane based analysis.

More recently, spatiotemporal processing has been used for periodicity analysis [24], camera work analysis [19], monitoring and surveillance applications [18] and motion analysis and segmentation [29, 37, 33]. Researchers have also used spatial

[46] and spatiotemporal salient point detectors [45] to enhance object recognition performance or region retrieval classification. Loccoz *et al.* in [45], go a bit further and attempt to combine spatiotemporal salient point detection with a higher level (knowledge-based) description of the underlying event. However, the majority of the proposed methods treat the video volume in spatiotemporal slices rather than in a volumetric manner. Spatiotemporal processing may solve several of the problems related to video analysis, but the large amount of data to be processed and the consequent computational burden may obstruct proper exploitation. Hence, a mechanism for selecting the meaningful part of the input to be processed is indispensable for designing successful, computationally efficient algorithms in the promising spatiotemporal domain. This need fits well with the saliency-based VA model discussed before. Related approaches that process only Regions-Of-Interest (ROI) are commonly used in video encoding, where certain parts of the visual input are of higher importance than the rest. Such ROI can be decoded with higher quality than the background or other non-important areas. There is already undergoing work that relates VA to JPEG 2000 and MPEG-4 standards [5, 38]. Additionally, recent work on video encoding demonstrates the improvement of the coding efficiency obtained by allowing variable bit allocation at the object level in both spatial and temporal domain [22, 9].

The proposed model is inspired by the *bottom-up saliency-based* VA, which has been computationally modeled in the last decade, [21, 15, 14, 30], and seems to provide a reasonable first step towards the elucidation and understanding of the visual input. This VA model was originally proposed by Koch and Ullman [21] and later implemented by Itti *et al.* [15]. In the model, all feature maps feed, in a purely bottom-up manner, into a master *saliency map*. The purpose of the saliency map is to

combine the “salient” locations from each of the lower feature maps (e.g., intensity, color, orientation, etc.) into a global measure weighting how much different a given location is from its surroundings. This computational strategy has been proven successful on many real images by providing robustness to noise and clutter. Although biologically inspired, the VA process has two lifelines along which its success might be measured as Tsotsos *et al.* [17] assert: *“The first is dependent on whether the biological predictions can be verified and whether new observations might be explained well by the model. The second is dependent on whether the model is useful in computational solutions of vision”*. We focus on extending the saliency-based VA in order to obtain an efficient computational vision model for video analysis.

In an attempt to build a general framework that will provide a reliable way to exploit both the static and dynamic (temporal) information of a video, we extend the frame-based computational VA model of Itti *et al.*’s [15] and treat the temporal dimension of a sequence as an intrinsic part of it [20]. In the current paper we provide an extensive step-by-step overview of the model and a concrete set of experiments to evaluate its performance. We treat the video sequence as a video volume with temporal evolution being the third dimension. Consequently, the movement of an object can be regarded as a volume carved out from the 3D space. Simple 3D operators along with morphological tools are used for extracting and enhancing features of interest. Under such a framework, locating and analyzing interesting events in a sequence by considering their actual temporal evolution across a large number of frames could be done without the need for, e.g., optical flow estimation, which is computationally expensive and possibly inaccurate, due to the small number of processed frames.

The proposed model combines the advantages of both VA and spatiotemporal processing and qualifies as a platform for various applications that involve event detection in noisy or occluded environments, surveillance and monitoring or as a preprocessing step for scene segmentation. The spatiotemporal VA can also serve as a tool for perceptual coding, since saliency encoded in the output of the system can be used by a rate control mechanism to improve subjective quality and highlight the importance of content interpretation.

The paper is organized as follows: Section 2 introduces the proposed spatiotemporal visual attention scheme and provides an in-depth overview of the model and its main components. In Section 3 experimental results are given, while future work is discussed and conclusions are drawn in section 4.

2. BOTTOM-UP SPATIOTEMPORAL VISUAL ATTENTION

Fig. 1 provides an illustrative view of the approach. The input video is decomposed into feature volumes, which encode intensity, colors and orientations. Separate conspicuous volumes are generated and finally combined to produce the saliency volume. The proposed model consists of several intermediate steps, which are shown in Fig.2. The whole procedure may be divided into video preprocessing, feature volume generation and saliency volume generation. The following sections give a detailed overview.

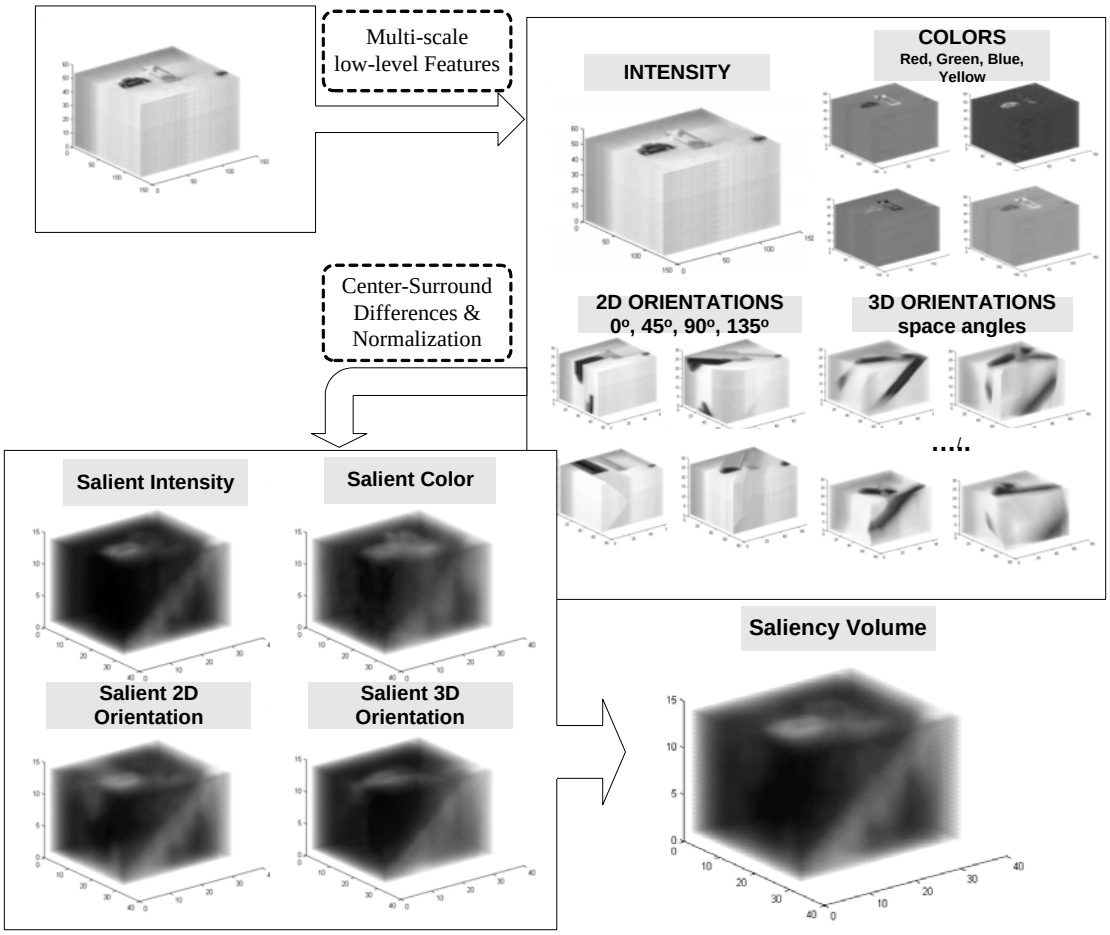


Figure 1 Spatiotemporal VA architecture. The Feature Extraction stage and the Saliency Volume generation are shown.

2.1 Video Preprocessing

2.1.1 Video Volume Generation

Given an arbitrary input sequence, the first processing step consists of temporally segmenting the sequence into a set of video shots using a common shot-detection technique [32]. The number of frames to be processed with the proposed computational model can be the same as the length of the corresponding shot, or a number of frames that is sufficient to represent adequately the objects' trajectories across them. In other words, we consider a spatiotemporal block of frames of a video sequence that are (relatively) closely sampled. Hence, we treat the video sequence as

a volume with (*x, y*) being the spatial dimensions and *t* (frame evolution) the temporal one. Specifically, the spatial dimensions of width and height are the *x*- and *y*- axes of a frame, while the temporal one is derived by layering the frames sequentially in time (*x*-*y*-*t* space). The minor element of a volumetric representation is called voxel and is denoted throughout this paper as *v*. Such a volumetric representation provides richer structure and organization along a large temporal scale than individual 2D frames and leads to more complete video analysis approaches in terms of data exploitation.

The video volume is decomposed into a set of distinct "channels" by using linear filters tuned to specific stimulus dimensions, such as intensity and red, green, blue hues. The number and response properties of these filters have been chosen according to what is known of their neuronal equivalents in the early stages of visual processing in primates [15]. Hence, the red channel volume, for example, denoted by $r(x, y, t)$ (or simply *r*) is characterized by the spatiotemporal data spanning the three dimensions as explained above. The green and blue channel volumes are similarly denoted as *g* and *b*, respectively. The intensity volume is then obtained as $I = (r + g + b)/3$. Every volume simultaneously represents the spatial distribution and temporal evolution of the encoded feature. Interestingly enough, by exploiting the last consideration, we avoid the motion estimation needed in other proposed methods to infer the dynamic nature of the video content.

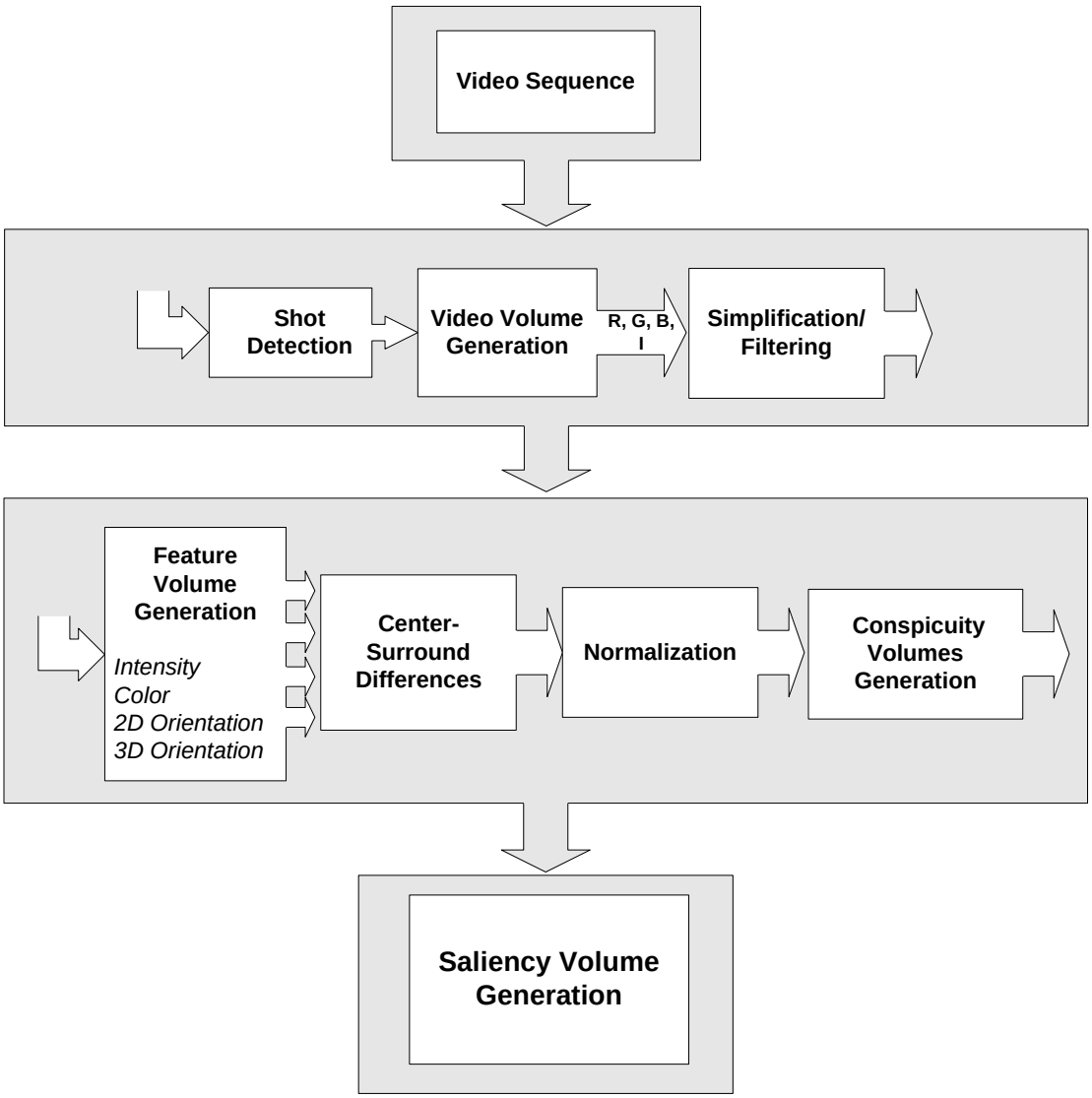


Figure 2 Spatiotemporal VA model

2.1.2 Simplification/Filtering

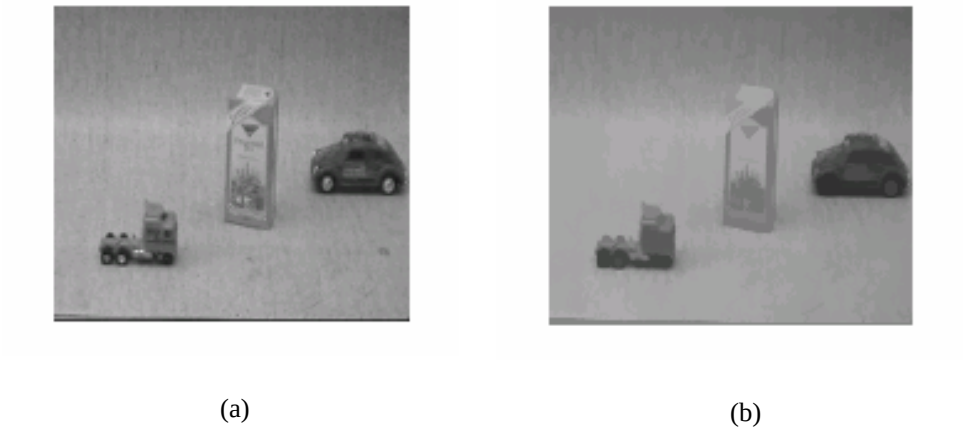
After obtaining spatiotemporal data formation, the input volumes are filtered so as to avoid spurious details or noisy areas that might otherwise be erroneously attended by the proposed system. The main objectives of this filtering stage are noise removal and simplification of intensity/color components. Nevertheless, the simplification algorithm should definitely retain the edge structure and produce homogenous areas between edges.

Morphological connected operators fit well with this task. We employ connected filters, in particular those called *filters by reconstruction*, because of their attractive property of simplifying the image while preserving contours. The flat-zones are computed by the use of *Alternating Sequential Filters* (ASF), which are based on morphological area opening and closing operations with structuring elements of increasing scale [7][27][36]. Particularly, if *nS* is a 3D structuring element of increasing size n=1,2,3,… then the openings α_n and closings β_n that make up the filter for a volume V, are

$$\begin{aligned} \alpha_n(V)(x,y) &= \max\left[I \circ nS(x,y,t)\right] \\ \beta_n(V)(x,y) &= \min\left[I \bullet nS(x,y,t)\right] \end{aligned} \tag{1}$$

where V may be any of the r, g, b, I volumes. Finally, the filtered volume V_{ASF} is obtained by $V_{ASF} = \beta_n \alpha_n ... \beta_2 \alpha_2 \beta_1 \alpha_1$. Fig. 3 illustrates an example of the application of the *ASF* filter on a two-dimensional image.

The simplification procedure using ASF is applied to each of the four main channels, namely intensity, red, green and blue channels, while the yellow one is a combination of the other three.



(a) (b)

Figure 3 Illustration of image simplification using the flat-zones approach: (a) original, (b) simplified frame

2.2 FEATURE VOLUME GENERATION

Following the structure of the static image-based approach of Itti & Koch, we generate pyramids of volumes for each feature of interest, including intensity, color, and orientation. Each of them encodes a certain property of the video. The different scales are created using Gaussian pyramids (Burt *et al.* [6]), which consist of progressively low-pass filtering and subsampling the input. In our implementation, the depth of the pyramid depends on the input video spatiotemporal size, but cannot be less than 5 scales. Low-pass filtering and subsampling is obtained by 3D Gaussian low-pass filters and vertical/horizontal reduction by consecutive powers of two. The final result is a hierarchy of video volumes that represent the input in decreasing spatiotemporal scales. This decomposition is done for each of the feature discussed in the following subsections and allows the model to represent smaller and larger "events" in separate subdivisions of the channels.

2.2.1 Intensity and Color

A Gaussian pyramid of intensity volumes $\mathbf{I}(\sigma)$ is created, where $\sigma=1,\ldots,5$ is the pyramid scale (level). Color channels are normalized by I in order to decouple hue from intensity. Since hue variations are not perceivable at very low luminance, normalization is only applied to the locations where I is larger than $1/10$ of its maximum over the entire volume (other locations yield almost zero r, g, b). Four broadly tuned color channels (red, green, blue and yellow), suggested by the opponent color theory, are created [15]:

$$\begin{aligned} \mathbf{R} &= r-(g+b)/2, \quad \mathbf{G} = g-(r+b)/2 \\ \mathbf{B} &= b-(r+g)/2, \quad \mathbf{Y} = (r+g)/2-|r-g|/2 \end{aligned} \tag{2}$$

Each channel yields maximal response for the hue to which it is tuned and zero response for both black and white inputs. Four Gaussian spatiotemporal pyramids $\mathbf{R}(\sigma), \mathbf{G}(\sigma), \mathbf{B}(\sigma), \mathbf{Y}(\sigma)$ are created in this way.

2.2.2 Orientation

Spatial orientation can be calculated by extracting oriented edges at each frame and superimposing the results, while temporal orientation is obtained by direct 3D filtering of the video volume. 3D filtering is related to motion analysis tasks since orientation in space-time corresponds to velocity [10]. In order to get orientation, one needs an appropriate three-dimensional steerable filter set and a method to extract a measure of orientation out of the filters' output. Although motion is of fundamental importance in biological vision systems and contributes to visual attention, as confirmed by Watanabe *et al.* [39], it is not included as a feature map in the saliency-based computational model of Itti & Koch [15]. Elsewhere we have used motion for event analysis purposes [34]. Actually, motion/velocity description of the objects can be directly extracted by the 3D orientation volume as described in [33][12][40] avoiding therefore the need for independent optical flow computation.

2D Orientation

Gabor pyramid decomposition (or steerable filters decomposition [10]) is widely used for local orientation calculation in static images due to the widely accepted belief of an existing biological counterpart. Unfortunately, straightforward extension to spatiotemporal domain is computationally demanding, thus we use a different approach to obtain similar results in a computationally efficient way. Hence, we design an algorithm based on morphological tools that shares common ground with

the orientation module of Itti *et al.*'s prototype visual attention scheme. Orientation information is obtained from I using morphological processing of the corresponding Gaussian pyramids. Generally speaking, the morphological operations transform the original image into another one through the interaction with a structuring element of certain shape and size. Geometric features of the image that are similar in shape and size to the structuring element are preserved while other features are suppressed. Therefore, morphological operations can simplify the image data in a controlled way by preserving the desired characteristics and eliminating the irrelevant ones. We use oriented line structuring elements $\boldsymbol{S_\theta}$ in order to obtain the main 2D orientations at each frame of the video volume and generate the 2D orientation volume as

$$\mathbf{O}(\theta,\sigma) = \mathbf{I}(\sigma) \circ S_\theta, \quad \theta \in A = \left\{0^o, 45^o, 90^o, 135^o\right\} \tag{3}$$

Applying an opening to the input frame with an oriented structuring element of specific angle yields strong response at the corresponding angle. The opening is applied to each level of the pyramid with oriented structuring elements of decreasing lengths (power of two).

3D Orientation

Spatiotemporal volumes can be seen as a composition of numerous simple structures like planes, textures, edges and lines. Therefore multiple oriented structures may be present at a single point. The volume can be either decomposed into images, as traditionally carried out, or into overlapping 3D local neighborhoods. Loosely, a neighborhood of voxel v is defined as the proximate voxels surrounding v. By using 3D connectivity, we can apply 3D morphological operations at every volume. We filter the volume with rotated versions of an orientation-selective morphological structuring element and produce a result with enhanced oriented subvolumes being

the result of the objects' path in the scene (3D). Hence the same procedure described in section 4.2.1 is applied to the spatiotemporal domain. Cylinder-shaped structuring elements are used in order to obtain the desired 3D orientations of the video volume:

$$\mathbf{O}_{3D}(\theta,\sigma) = \mathbf{I}(\sigma) \circ S_{\theta}^{3D}, \quad \theta \in A^{3D} \tag{4}$$

Five of the nine main orientations of A^{3D} are illustrated in Fig. 4.

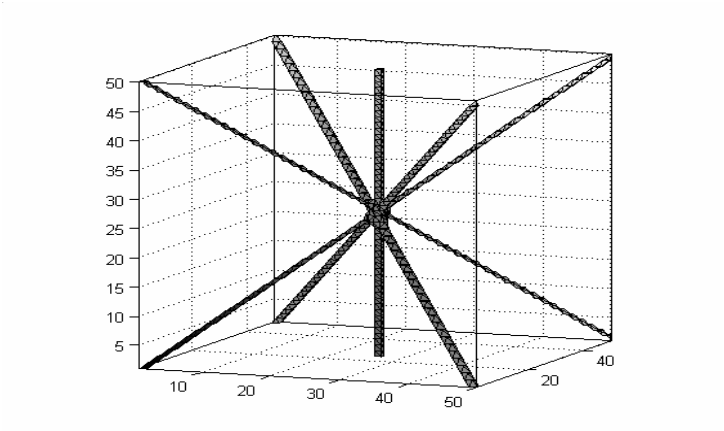


Figure 4 Five of the nine main 3D orientations (not all of them are shown for illustrative purposes)

2.3 SALIENCY VOLUME GENERATION

2.3.1 Center-Surround Differences

Let us first concentrate on how retina neurons operate. The area of the retina within which a neuron's activity can be influenced is referred to as that neuron's *receptive field*. Because each part of the retina corresponds to a region of visual space, a receptive field can also be defined as the part of visual space to which a neuron responds. There are several different types of retinal ganglion cell receptive fields. By far the most common type is the center-surround receptive field. These receptive fields are characterized by circular symmetry, and the presence of two distinct,

mutually-antagonistic sub-regions, a center and a surround. The sensitivity profiles of the centers and the surrounds are thought to be Gaussian: the response of the center is smaller near its border than at its midpoint. Such architecture is well-suited to detecting locations that locally stand-out from their surround. Center-surround, denoted as $\ominus$, is implemented in the model as the difference between fine and coarse scales. The center is a pixel at scale $c \in \{2,3\}$, and the surround is the corresponding pixel at scale $s = c + \delta$, with $\delta \in \{1,2\}$. Hence, we generate five feature volumes as follows:

$$\mathbf{I}(c,s) = |\mathbf{I}(c) \ominus \mathbf{I}(s)| \tag{5}$$

$$\mathbf{RG}(c,s) = |(\mathbf{R}(c) - \mathbf{G}(c)) \ominus (\mathbf{G}(s) - \mathbf{R}(s))| \tag{6}$$

$$\mathbf{BY}(c,s) = |(\mathbf{B}(c) - \mathbf{Y}(c)) \ominus (\mathbf{Y}(s) - \mathbf{B}(s))| \tag{7}$$

$$\mathbf{O}(c,s) = |\mathbf{O}(\theta,c) \ominus \mathbf{O}(\theta,s)| \tag{8}$$

$$\mathbf{O_{3D}}(c,s) = |\mathbf{O_{3D}}(\theta,c) \ominus \mathbf{O_{3D}}(\theta,s)| \tag{9}$$

2.3.2 Normalization

There is an intrinsic difficulty in combining all the volumes resulting from the feature extraction stage. When no knowledge about the scene exists, there is no way to bias the systems towards specific (salient) features. The spatiotemporal feature volumes represent *a priori* not comparable modalities, with different dynamic ranges and meaning. Due to the lack of top-down supervision (knowledge), there is a need for a

normalization scheme that will enhance high activation areas and suppress others. Such a scheme will enhance the most salient subvolumes so as to prohibit non-salient regions from drastically affecting the result. We use a simple normalization operator N that consists of the following: 1) normalize all the spatiotemporal feature volumes to the same dynamic range, in order to eliminate across-modality amplitude differences; 2) for each volume find the global maximum M and the average $\overline{m}$ over all other local maxima; 3) globally multiply the volume by $(M-\overline{m})^2$.

Emphasizing on morphological approaches, which are computationally less demanding and operate in a controlled way, we use the grayscale top-hat transformation ($THT(V,S)=V-(V\circ S)$) for obtaining the local maxima of each volume. Comparing the maximum activity area to the average over all other maxima areas measures how important, in terms of intensity value, is the most active area. When this difference is large, we strongly promote the map.

2.3.3 Conspicuity and Saliency Volumes Generation

The feature volumes are combined into four *conspicuity volumes*, $\bar{I}$ for intensity, $\overline{C}$ for color, $\overline{O}$ for 2D orientation and $\overline{O}_{3D}$ for 3D orientation at an intermediate scale (σ_i) of the spatiotemporal decomposition. They are obtained through across-scale addition, $\oplus$, which consists of reduction of each volume to scale σ_i and point-by-point addition ([15]):

$$\bar{I}=\bigoplus_{c=2}^{\sigma_i}\bigoplus_{s=c+1}^{\sigma_i}N(\mathbf{I}(c,s)) \tag{10}$$

$$\overline{C}=\bigoplus_{c=2}^{\sigma_i}\bigoplus_{s=c+1}^{\sigma_i}\left[N(\mathbf{RG}(c,s))+N(\mathbf{BY}(c,s))\right] \tag{11}$$

$$\overline{O}=\sum_{\theta\in A}N\left(\bigoplus_{c=2}^{\sigma_i}\bigoplus_{s=c+1}^{\sigma_i}N(\mathbf{O}(c,s))\right) \tag{12}$$

$$\overline{O}_{3D} = \sum_{\theta \in A^{3D}} N\left(\bigoplus_{c=2}^{\sigma_i} \bigoplus_{s=c+1}^{\sigma_i} N\left(\mathbf{O_{3D}}(c,s)\right) \right) \tag{13}$$

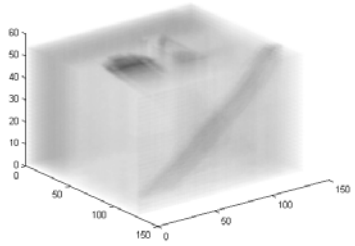
The motivation for the creation of the separate channels and their individual normalization is the hypothesis that similar features compete strongly for saliency, while different modalities contribute independently to the saliency volume. Finally, the four conspicuity volumes are normalized and summed into the saliency volume

$$S = \frac{1}{4}\left(N(\bar{I}) + N(\overline{C}) + N(\overline{O}) + N(\overline{O}_{3D})\right) \tag{14}$$

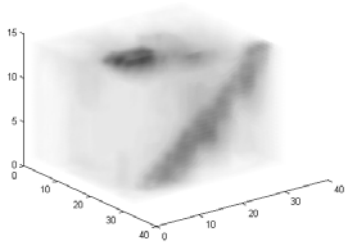
In order to illustrate the three-dimensional aspect involved in the proposed architecture we show representative views of the saliency volume obtained from a simple sequence acquired by a static camera. The "*truck*" sequence shows two toy-trucks moving towards opposite directions. A static box in the middle of the scene occludes one of them. Fig. 5a-b show three representative frames of the sequence and the semi-transparent volume of the original sequence and three representative frames, while Fig. 5c-e show the saliency volume under three different angles. All of them are negative and transparent versions of the original saliency volume (for visualization purposes). The route of the first truck, which is visible throughout the sequence, is highlighted as a consistent black cylinder at the top-right volume. The temporal evolution of both moving trucks is shown clearly at the bottom-left image, while the vertical pattern generated by the static box is illustrated at the bottom-right subfigure.

3 RESULTS

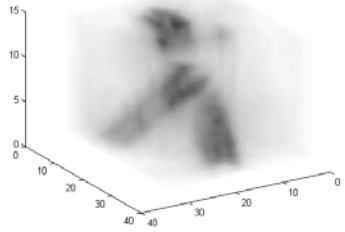
(a)



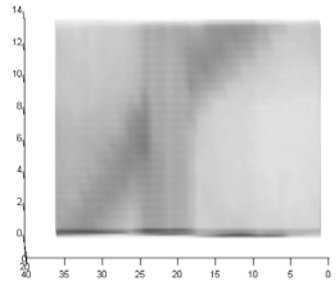
(b)



(c)



(d)



(e)

Figure 5 (a) three representative frames of the sequence; (b) unprocessed video volume; (c)-(e) saliency volume observed from 3 different angles. The volumes are negative and transparent versions of the original saliency volume for visualization purposes.

3.1 Experimental Setup

Illustrating the power of the proposed spatiotemporal VA architecture is not easy due to the three dimensional data and the inherent visualization problems. Hence, we present the results by using a see-through mask for every input frame that is directly acquired from the corresponding *x-y* slice of the saliency volume. Specifically, the saliency volume of a sequence looks like the one illustrated at Fig. 5a. The intensity of each voxel is related to the saliency of that pixel.

For visualization purposes, we interpolate the volume and produce one with the same dimensions as the input sequence, Fig. 6b. Slicing this volume across the temporal dimensions at every time frame produces a saliency map for each of the input frames Fig. 6c. Superimposing this mask on the corresponding frame generates

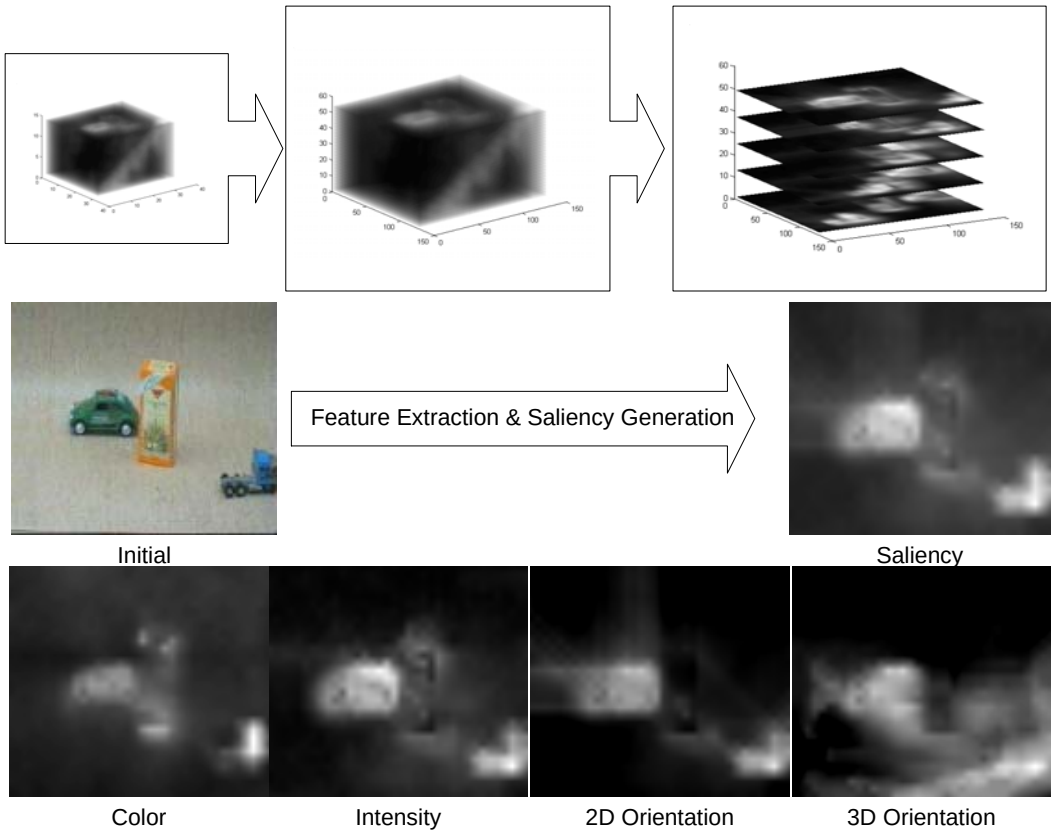


Figure 6 (a)-(c) Generation of mask (see text) for visualization purposes; (d) the initial frame and corresponding slices for each of the feature and saliency volumes.

the desired result. Non-salient areas appear dark, while salient ones preserve (almost thoroughly) their original intensity. It is important to mention that no thresholding is applied to the final masks. Fig. 6d shows an initial frame of the "*truck*" sequence and corresponding slices for each of the feature and saliency volumes.

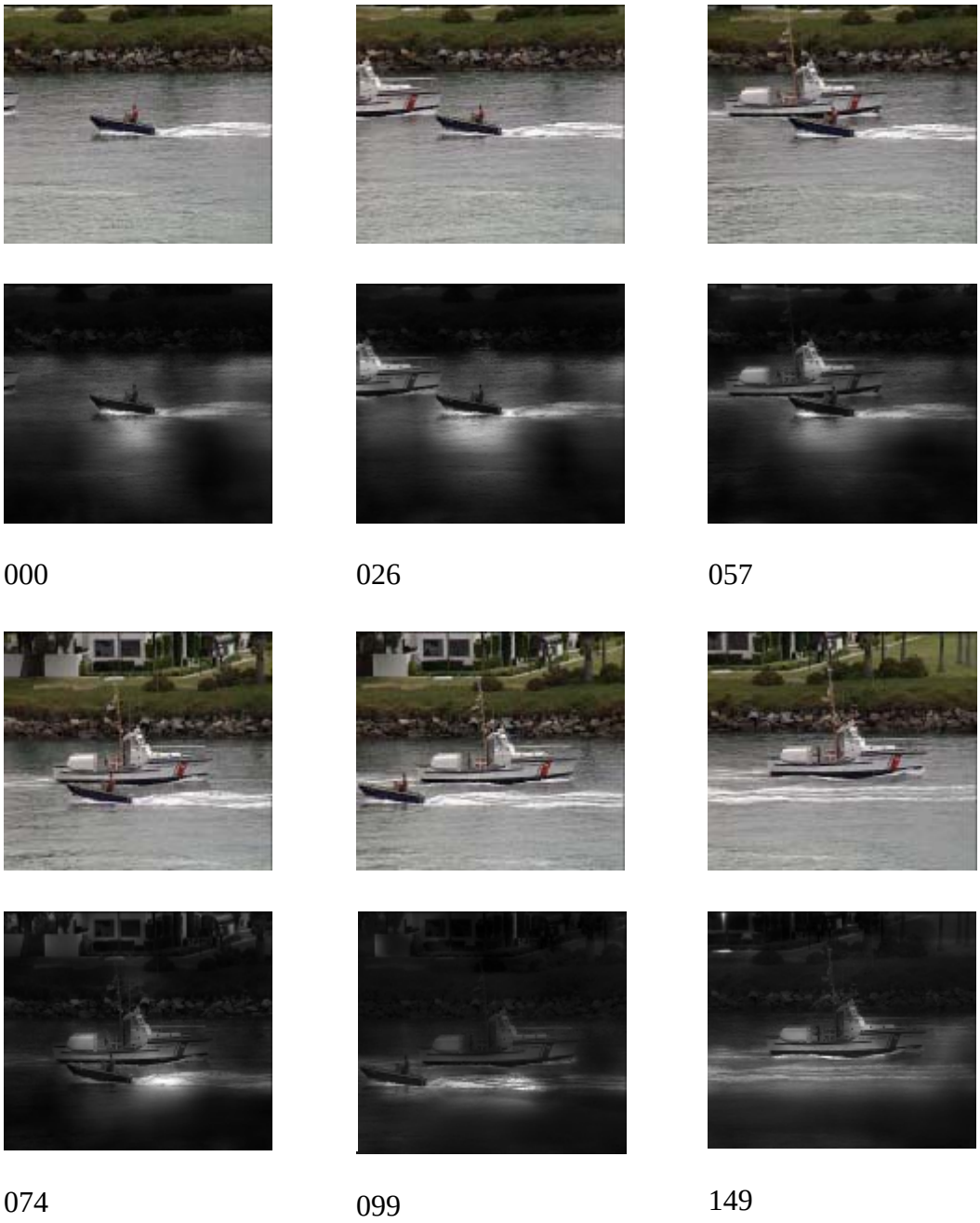


Figure 5 Results on the "coast-guard" sequence (numbers correspond to frames). Notice that the global camera motion does not affect the result.

To illustrate the behaviour of the model, we first consider natural image sequences and provide a qualitative analysis of the results. The sequences analyzed in section 3.2 are in QCIF format (176x144) and are processed in chunks of 100-150 frames (shot's length dependent). The quantitative analysis presented in section 3.3 uses ground-truth data from the CAVIAR dataset, [41], to compare the proposed method with the established saliency-based approach of Itti *et al.* [14]. The resolution of CAVIAR data

is half-resolution PAL standard (384 x 288 pixels, 25 frames per second) and the ground-truth was obtained by hand-labeling the images. In order to be fair, we used the publicly available Neuromorphic toolkit, implemented by Itti and his colleagues [44], for generating saliency maps for the Itti *et al.*'s method. These saliency maps are obtained using color, intensity, orientation and motion as feature maps.

3.2 Spatiotemporal saliency detection

Obtaining a successful segmentation of a dynamic scene is an important task for visual understanding. Most of the current video segmentation methods use e.g. intensity or color cues and exploit motion information either as the dominant or as supporting criterion for distinguishing between salient and non-salient objects/regions. Spatiotemporal VA, when used as a preprocessing step, may aid the accurate segmentation step that follows. The integration of several features under the proposed framework captures all the relevant object information and avoids the computationally demanding motion estimation step, since large uniformly moving regions that do not differ from their spatiotemporal neighborhood are not attended. Such regions may e.g. belong to the background that undergoes the same motion due to a moving camera. We illustrate this fact by showing the results on two well-known sequences, namely "coast-guard" and "table tennis".

The *"coast guard"* sequence shows a complex scene with different objects present. Two boats are moving in opposite directions in a river, while the camera pans, following the smaller boat initially and then the larger one. Trees and rocks cover the coast and the river presents wavy patterns throughout the sequence. The relative motions of the small and the large boats during the first and second pans are small in

magnitude due to the simultaneous movement of the camera. The proposed VA system performs well, since it “focuses” on the two boats and their
immediate surroundings without being affected by the camera motion or the minor changes on the river and the coast. It has to be mentioned that not only moving objects are of interest when preprocessing a video sequence. The proposed VA model focuses mainly at the two moving boats throughout the sequence, but it also enhances the saliency of the surge of the river around frame 099 as can be seen in Fig. 9. The surge becomes salient due to its spatiotemporal difference from the surroundings. After few frames (frame 149) the large boat becomes salient again.

The *“table tennis”* sequence presents a whole range of situations that makes it a challenging stream. Many of the regions of interest are discontinuous and rapidly changing. An interesting part of the sequence is the zooming out effect appearing approximately after the first 25 frames. The camera zooms out, but remains focused on a region between the ball and the bat. The challenge is to consistently distinguish the ROIs without being affected by the camera motion (zoom out). The first two columns of Fig. 10 show the original frame and the corresponding saliency mask derived from the saliency volume as explained above. The spatiotemporal VA system focuses at the player and the poster on the left even during the camera zoom-out (frames 25-86). Consistent distinction of the player and the incoming poster from the left can be achieved without being affected by camera operations as observed throughout the sequence.

Several proposed tracking techniques use motion information as an automatic initial guess for object’s position or for improving an incremental tracking approach [11, 25, 28]. Generally speaking, motion estimation methods are computationally intensive and prone to noise. Although the moving objects in a sequence are usually

important and have to be tracked there are cases that static objects (e.g. a scene with a camera pan showing a moving object and a photo/painting on the wall) play also a considerable role. In an attempt to emphasize the power of VA as a preprocessing step we provide a short discussion on the spatiotemporal VA's results of the "*table tennis*" sequence and the advantages it offers against a robust motion estimation technique that is used as an initial step in a tracking approach we proposed in [11]. The magnitude of the motion field generated by Black & Anandan's method [26] is shown

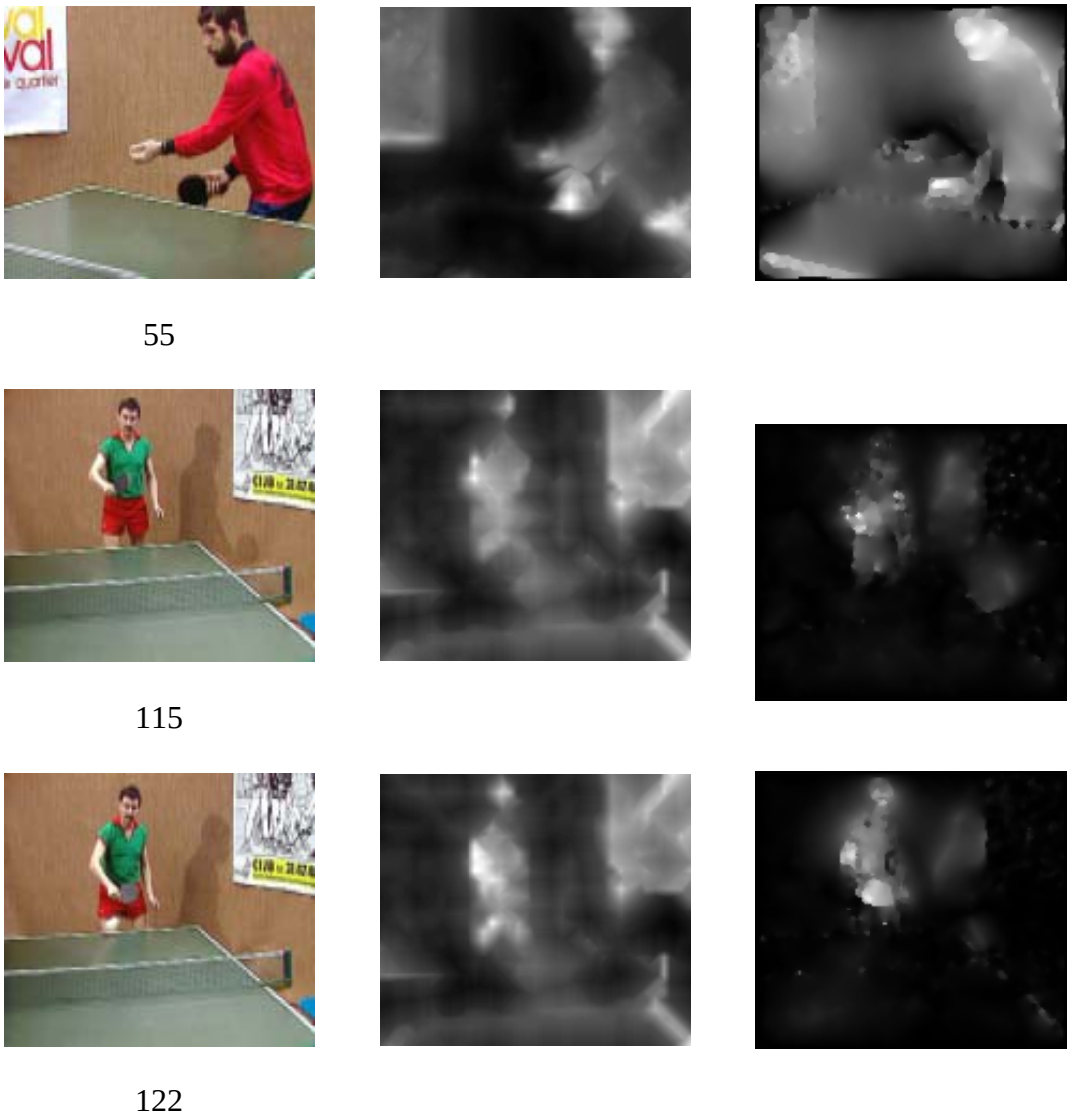


Figure 6 Results on the "table-tennis" sequence (numbers correspond to frames). Row-wise: original frame, saliency map and magnitude of the motion map

in the third column of Fig. 10. Notice how the zooming effect (frames 55, 75) affects the motion field and how hard it is to automatically distinguish the objects even with a refined motion segmentation technique. Spatiotemporal VA focuses on the salient objects (player, poster) without being affected by the overall change of the scene. The rest of the frames illustrate the ability of the VA to focus on objects that do not differ in terms of motion from the background. The motion estimation result can be correctly used for locating and tracking the player, but it provides no information on the poster at the right. Hence, the proposed spatiotemporal VA provides a richer representation of the scene, in terms of salient regions, that can aid a refined segmentation based on low-level (feature volumes) or high-level (e.g. knowledge about the relative position of static-dynamic objects etc.) information.

3.3 Judging spatiotemporal saliency

Validating the performance of saliency-based techniques in real computer vision problems is not a straight-forward task due to lack of appropriate ground-truth. Nevertheless, the efficiency of the visual attention model proposed by Itti *et al.* has been proven through experiments that compare human eye fixations with the ones obtained by the model [14][15]. In order to obtain numerical statistics that validate the methods' performance we set up an experiment that compares the outputs of the proposed spatiotemporal VA and Itti's one using part of the publicly available CAVIAR data [41]. The videos we used have been captured with wide angle lens along and across the hallway in a shopping centre in Lisbon and in the entrance lobby of the INRIA labs at Grenoble, France. Ground truth for these sequences was obtained by hand-labeling the images using filled bounding rectangles. Indicative frames of

five different videos and the corresponding ground-truth are shown in the first and last columns of Figs. 8, 9 .

In order to obtain a simple but robust segmentation of the saliency map we use a series of image analysis operators as shown in Fig. 7. The initial saliency map is first thresholded using the Otsu's method [42] and then morphologically filtered to reduce artifacts and fill the holes (Fig. 7c). Notice that the automatic thresholding does not introduce any undesired artifacts since the salient and non-salient areas are clearly separated in the initial map. Afterwards, a marker for each candidate region is obtained by dilating the regional maxima of the distance transform (Fig. 7e). The watershed, constrained by the markers image is then applied to the negative image of the distance function (Fig. 7g). The result of this procedure is also called geodesic SKIZ (Skeleton Influence Zone) [43]. The final mask is obtained by intersecting the filled image (Fig. 7c) with the negative version of the detected watersheds (Fig. 7g).

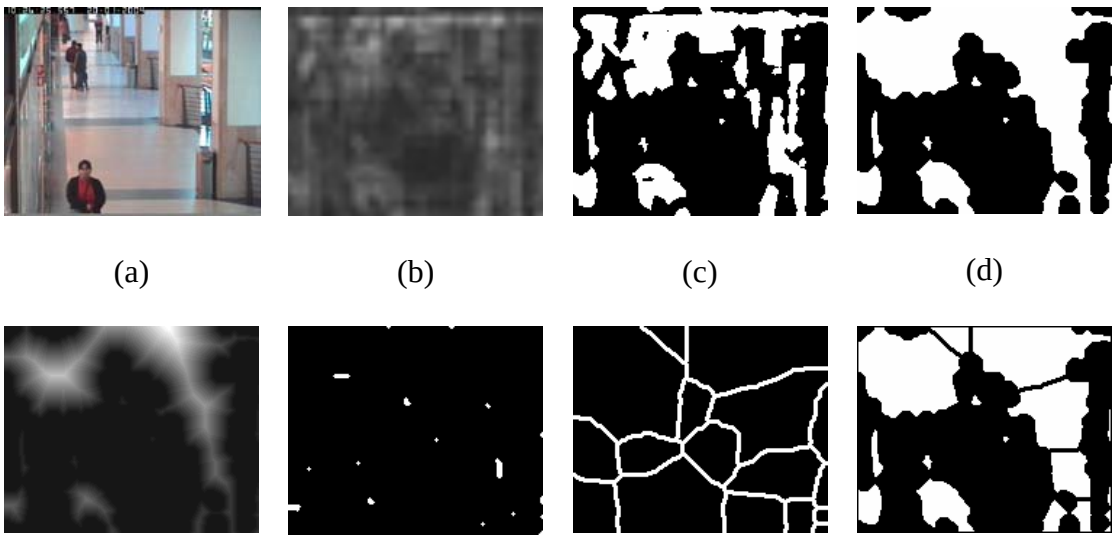


Fig. 7 Segmentation of a single saliency map; (a) initial frame; (b) saliency map; (c) Otsu thresholding; (d) morphological filtering; (e) distance transform; (f) regional maxima; (g) constrained watershed transform using (f) & neg(e); (h) intersection of (c) & neg(g)

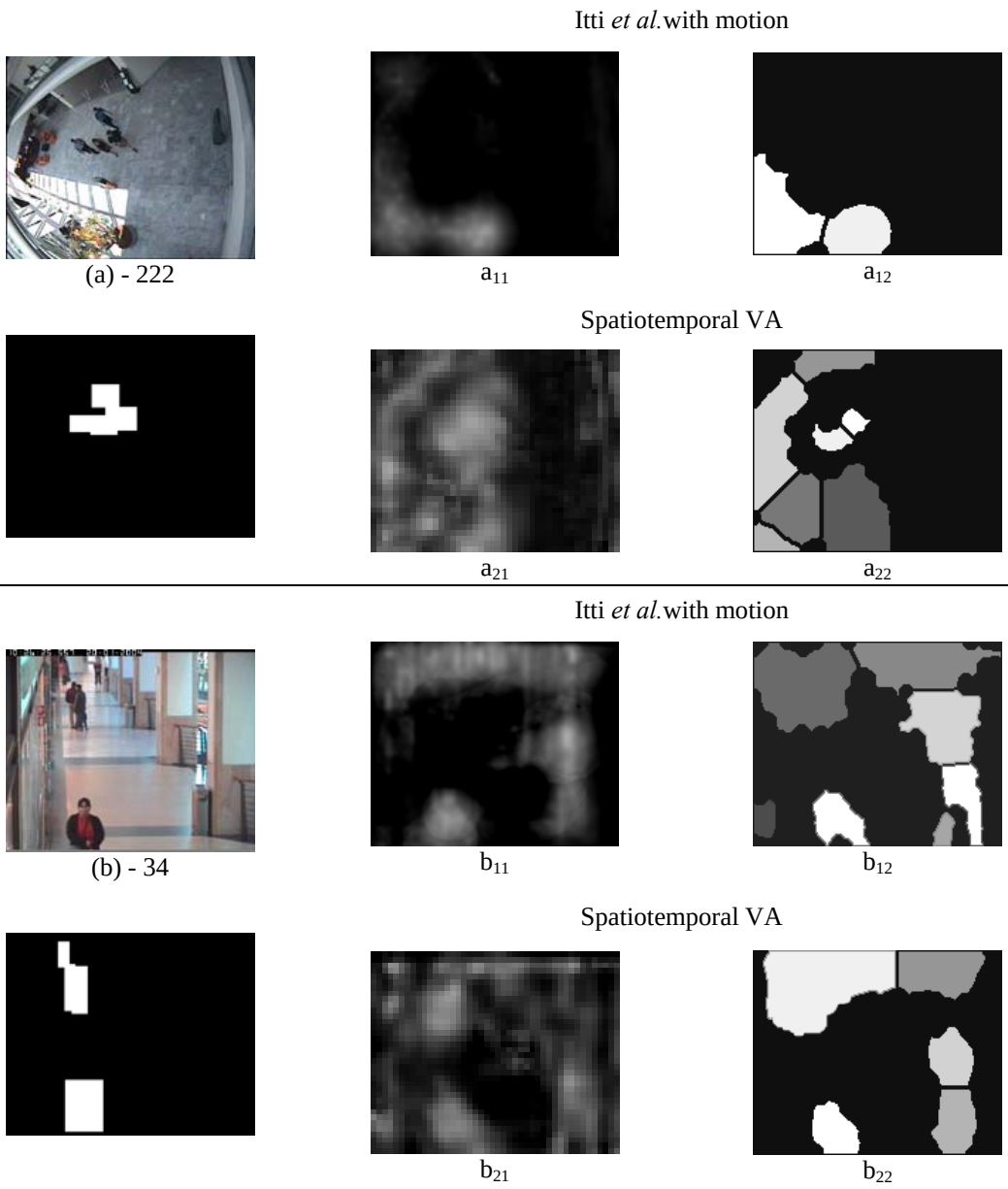


Fig. 8 Example frames of two different sequences along with the corresponding ground-truth found in CAVIAR data set (in row-wise order).

Indicative results for different videos of the CAVIAR dataset are shown in Figs. 8, 9. The first and last columns contain the original frame and the corresponding ground-truth, while the second column shows the saliency map obtained using Itti's method and the proposed spatiotemporal one. The segmentation result is shown at the third column. The separated regions are labeled according to saliency and filled in with different gray levels. Light gray regions correspond to high-salient ones, while darker regions are non-salient. As mentioned before, the ground-truth is composed of solid bounding boxes of people walking around a hall or along corridors. Visual attention

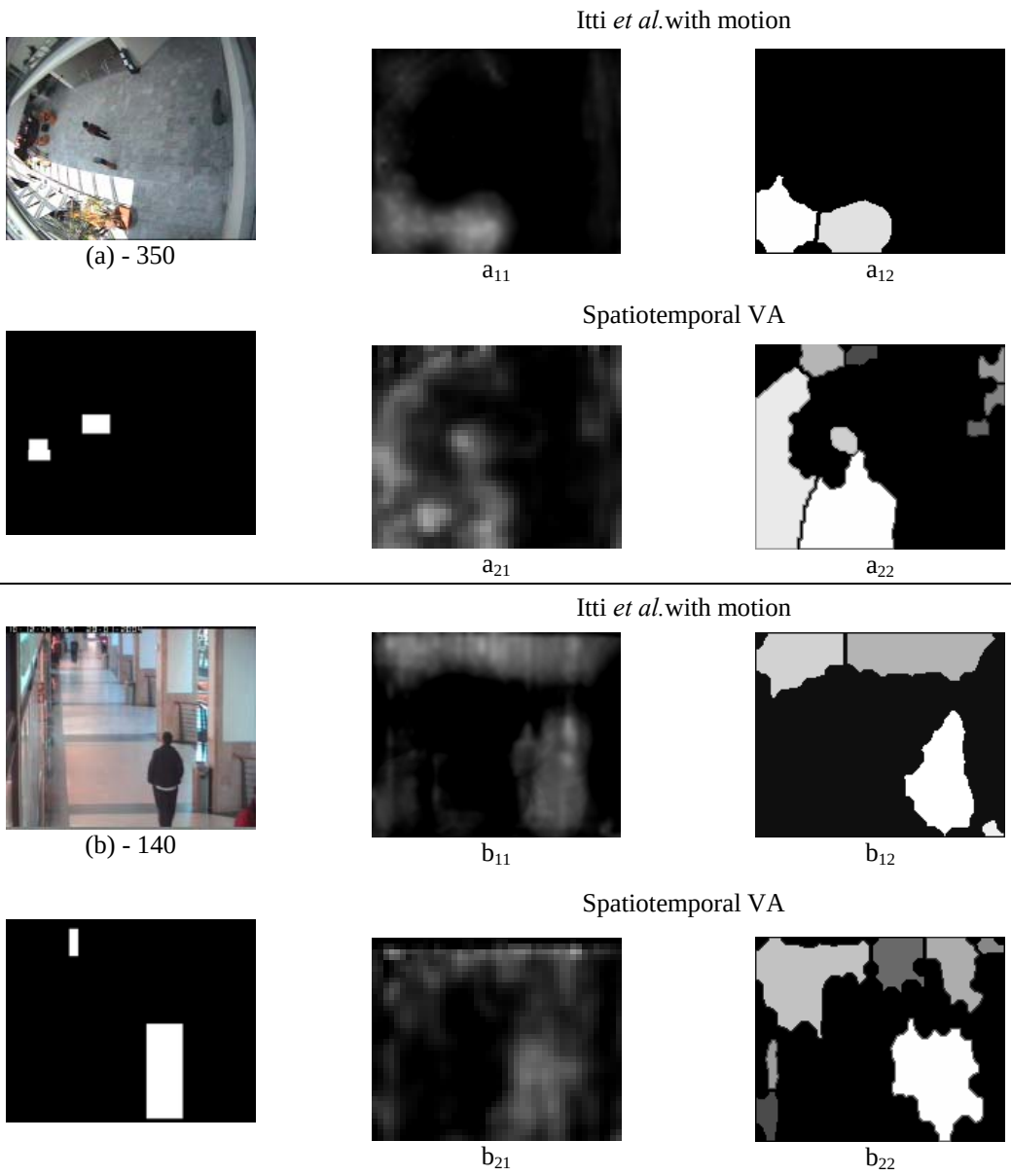


Fig. 9 Example frames of two different sequences along with the corresponding ground-truth found in CAVIAR data set (in row-wise order).

methods do not focus on regions that outline certain objects, but regions that probably attract the human eye and possibly contain an object of interest. Hence, statistical judgment of VA methods in real scenarios by comparing directly the segmented results with accurate masks of objects of interest (available ground-truth) is not fair. Additionally, ground-truth images usually contain objects having specific features or behaviour, like the ones found in CAVIAR datasets (walking people) making this comparison even more unfair.

Nevertheless, a VA technique should drag the focus-of-attention at these objects during time evolution. Under this framework, we establish two evaluation metrics: the absolute recall/precision values that the method can reach and the number of foci needed to reach these values or equivalently, the number of foci needed to find whole or part of the desired target. It is clear that high precision is much harder to obtain than high recall. The segmentation technique we use is simple and aims to provide crude salient regions around possible targets and therefore high recall is what we expect. Actually, this is the main role of VA, namely to limit visual search into few and distinct areas of the input and aid the refined analysis steps that may follow.

In order to provide statistical results we measure the percentage of the targets' bounding boxes that VA covers after a number of sequential foci-of-attention. In our case the focus-of-attention has the shape of the segmented regions obtained as described before. For example, if we consider the segmented saliency map in Fig. $8a_{22}$, the first two foci-of-attention will be around the center (light-gray regions) and the rest will sequentially cover the left part of the frame.

We follow this procedure for hundreds of frames of the available videos and calculate precision as the average ratio of the common area covered by both masks (ours and ground truth) to the area of our masks and recall as the average ratio of the common area to the area of the ground truth masks. As discussed before, we do not expect high precision values for both methods, since accurate segmentation is not the focus of this work. We provide results for the kind of videos found in the dataset, namely the one capturing people walking around in a hall (e.g. Fig 8a, Fig. 9a) and the second one that captures people walking along corridors (e.g. Fig. 8b, Fig. 9b).

The procedure we follow is outlined into the following steps: 1) a spatiotemporal saliency cube is produced for every 64 frames of the input sequence; 2) the saliency

map of each frame is segmented using the procedure described before; 3) the segmented regions are ordered according to saliency and finally, 4) precision-recall statistics are calculated. The same procedure is applied for the Itti *et al.*'s model in a per frame basis. In order to be fair we generate a saliency cube using the saliency of each frame and filter it with a 3D median filter to enhance coherency.

The statistics plotted in Fig. 10 are the mean precision/recall on 3154 frames of the INRIA sequences, while the ones in Fig. 11 concern 3720 frames of the Lisbon sequences. The horizontal axis corresponds to the number of foci and the *y*- axis range is equal for the precision and recall plots respectively. Tables 1 and 2 outline the main statistical outcomes for both kinds of sequences and are interpreted as follows: e.g. for the proposed spatiotemporal VA the best recall value (71.57%) is obtained for the INRIA sequences (table 1) after 15 foci of attention, while the first focus-of-attention achieves a recall of 31.75%. Similarly, a precision of 14.20% is obtained after 8 foci-of-attention.

Table 1 Statistics on INRIA sequences

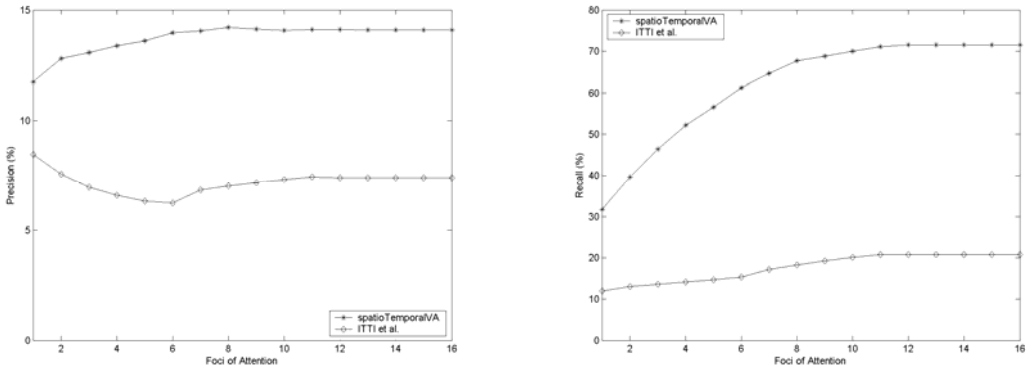
	Precision	**# foci**	**1st focus**	**Recall**	**# foci**	**1st focus**
Spatiotemporal VA	14.20%	8	11.75%	71.57%	15	31.75%
Itti *et al.* with motion	8.47%	1	8.47%	20.84%	14	11.96%

Table 2 Statistics on LISBON sequences

	Precision	**# foci**	**1st focus**	**Recall**	**# foci**	**1st focus**
Spatiotemporal VA	11.94%	2	11.55%	61.52%	15	35.90%
Itti *et al* .with motion	5.37%	1	5.37%	15.21%	14	11.21%

The mean precision/recall plots reveal the real strengths of the spatiotemporal VA technique. The proposed method clearly outperforms the second technique and reaches more than half of the max attained value at the first focus-of-attention. It is worth mentioning that the difference in performance between the two methods is

entirely due to the spatiotemporal configuration of the proposed model. Itti *et al.*'s method includes a motion channel and has been temporally median filtered for further temporal enhancement.



(a) (b)

Fig. 10 INRIA sequences (a) Precision; (b) Recall

The proposed configuration and the proposed 3D normalization/fusion operators capture temporal activities that differ from the surroundings in an efficient way and are not sensitive to minor, very short events which are likely to occur due to illumination flickering or shadows. INRIA sequences contain large targets in a relative non-complex background scene. LISBON sequences are more challenging since they contain small targets on a complex background (textured with bright sunlight). Statistics on both sequences are much higher for spatiotemporal VA than

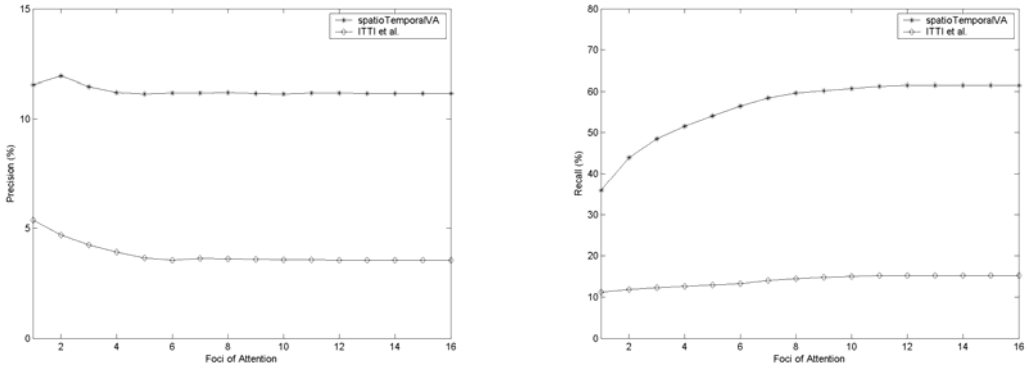


Fig. 11 LISBON sequences (a) Precision; (b) Recall

the method of Itti *et al.*

4 CONCLUSIONS & FUTURE WORK

A model that exploits spatiotemporal information for video analysis based on the concept of saliency-based visual attention has been presented. The goal is to provide an efficient pre-processing step (salient spatiotemporal event detection) that will limit the application of high level processing tasks to the most salient parts of the input.

Overall, the performance of the proposed model is very good and exploits both the spatial and temporal dimensions of the video. Both qualitative and quantitative evaluation of the model presented in this paper is quite promising. Fusing the proposed bottom-up spatiotemporal visual attention technique with prior knowledge (top-down) for putting up new applications in the field including classification, segmentation and tracking will be the focus of our future research.

3. RERERENCES

[1] Adelson E.H., Bergen J., "Spatiotemporal energy models for the perception of motion", J. Opt. Soc. Amer., vol. 2, no. 2, pp. 284-299, Feb. 1985.

[2] B. Horn and B. Shunck, "Determining optical flow", Artificial Intelligence, no. 17, 185–203, 1981.

[3] Bolles C.R., Baker H.H., "Generalizing epipolar-plane image analysis on the spatio-temporal surface", International J. Comput. Vision, vol. 3, pp. 33-49, 1989.

[4] Bolles C.R., Baker H.H., Mariont D.H., "Epipolar-plane image analysis: An approach to determining structure from motion", International J. Comput. Vision, vol. 1, pp. 7-55, 1987.

[5] Brandley P.A., Stentiford M.W., “Visual attention for region of interest coding in JPEG 2000”, J. Vis. Commun. Image R., vol. 14, pp. 232-250, 2003

[6] Burt P.J., Adelson E.H., “The Laplacian pyramid as a compact image code”, IEEE Trans. On Communications, vol. 31, pp. 532-540, 1983.

[7] Crespo J., Scaher W.R., Serra J., Gratin C., Meyer F., “The flat zone approach: A general low-level region merging segmentation method”, Signal Processing, vol. 62, pp. 37-60, 1997.

[8] E. Dickmanns, “Expectation-based dynamic scene understanding”, in (eds.) Blake & Yuille, Active Vision, MIT Press, Cambridge Massachusetts, pp. 303-334.

[9] F.C. Martins, W. Ding and E. Feig, “Joint control of spatial quantization and temproal sampling for very low bit rate video,” *Proc. ICASSP*, May 1997.

[10] Freeman W.T., Adelson E.H., “The Design and Use of Steerable Filters”, IEEE Trans. Patt. Anal. Mach. Intell., Vol 13 Num 9, pp 891-906, September 1991.

[11] G. Tsechpenakis, K. Rapantzikos, N. Tsapatsoulis and S. Kollias, “A snake model for object tracking in natural sequences”, Elsevier, Signal Processing: Image Communication, vol. 19, no. 3, pp. 219-238, Mar 2004.

[12] Huang C.L., Chen Y.T., “Motion estimation method using a 3d steerable filter”. *Image and Vision Computing*, vol. 13, pp. 21–32, 1995.

[13] Hubel D., “Eye, Brain, Vision”, New York: Scientific American Library, 1988.

[14] Itti L., Koch C., “A saliency-based search mechanism for overt and covert shifts of visual attention”, Vision Research, vol. 40, pp. 1489-1506, 2000.

[15] Itti L., Koch C., Niebur E., “A model of saliency-based visual attention for rapid scene analysis”, *IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI),* vol. 20, no. 11, pp. 1254-1259, 1998.

[16] J. L. Barron, D. J. Fleet, and S. S. Beauchemin, “Systems and experiment performance of optical flow techniques”, International Journal of Computer Vision, vol. 12, pp. 43–77, 1994

[17] J.K. Tsotsos, S.M. Culhane, W.Y.K. Wai, Y. Lai, N. Davis, F. Nuflo, “Modeling visual attention via selective tuning”, Artificial Intelligence, vol. 78, pp. 507-545, 1995.

[18] Jähne B., “Spatio-temporal image processing: Theory and scientific applications”, New York: Springer-Verlag, 1991.

[19] Joly P., Kim H.K., “Efficient automatic analysis of camera work and microsegmentation of video using spatiotemporal images”, Signal Process.: Image Commun., no. 8, pp. 295-307, 1996.

[20] K. Rapantzikos, N. Tsapatsoulis, and Y. Avrithis, “Spatiotemporal visual attention architecture for video analysis”, Proc. of IEEE International Workshop On Multimedia Signal Processing (MMSP 2004), 2004.

[21] Koch C., Ullman S., “Shifts in selective visual attention: towards the underlying neural circuitry”, Human Neurobiology, vol. 4, pp. 219-227, 1985.

[22] Lee, J-W.; Vetro, A.; Wang, Y.; Ho, Y-S., "Bit Allocation for MPEG-4 Video Coding with Spatio-Temporal Tradeoffs", *IEEE Transactions on Circuits and Systems for Video Technology*, ISSN: 1051-8215, Vol. 13, Issue 6, pp. 488-502, June 2003

[23] Leventhal A.G., “The neural basis of visual function”, Vision and Visual Dysfunction, vol. 4, Boca Raton, FL:CRC Press, 1991.

[24] Liu F., Picard R.W., “Finding periodicity in space and time”, in Proc. IEEE Int. Conf. On Computer Vision, pp. 376-383, 1998.

[25] M. Pardàs, E. Sayrol. “Motion estimation based tracking of active contours”, *Pattern Recognition Letters* , 22:1447-1456, 2001.

[26] M.J. Black, P. Anandan, “The robust estimation of multiple motions: parametric and piecewise-smooth flow fields”, CVIU, vol. 63, no. 1, pp. 75–104, 1996

[27] Maragos P., “Noise Suppression", The Digital Signal Processing Handbook, V.K Madisetti and D.B Williams Eds., CRC Press, Chapt. 74, pp. 20-21, 1998.

[28] N. Peterfreund, “Robust tracking of position and velocity with Kalman snakes”, IEEE Transactions on Pattern Analysis and Machine Intelligence, 21 (6): 564-569, 1999.

[29] Ngo C.-W., Pong T.-C., Zhang H.-J., “Motion analysis and segmentation through spatio-temporal slices processing”, IEEE Trans. On Image Processing, vol. 12, no. 3, Mar 2003.

[30] Niebur, E. and Koch, C., “Computational architectures for attention” In Parasuraman, R.,editor, *The Attentive Brain*, chapter 9, pages 163–186. MIT Press, Cambridge, MA., 1998.

[31] Pashler H. Attention and performance. Ann. Rev. Psych, vol. 52, pp. 629-651, 2001.

[32] Patel N.V., Sethi I.K., “Video Shot Detection and Characterization for Video Databases”, Pattern Recognition, vol. 30, no. 4, pp. 583-592, April 1997.

[33] Porikli F., Wang Y., “Automatic video object segmentation using volume growing and hierarchical clustering”, EURASIP Journal on Applied Signal Processing (Object-based and Semantic Image & Video Analysis), vol. 2004, no. 6, pp. 814-832, Jun 2004.

[34] Rapantzikos K., Tsapatsoulis N., “On the implementation of visual attention architectures”, Tales of the Disappearing Computer, Santorini, June 2003.

[35] S. Baluja, D.A. Pomerleau, “Expectation-based selective attention for the visual monitoring and control of a robot vehicle”, Robotics and Autonomous Systems Journalm vol. 22, pp. 329-344, 1997.

[36] Salembier P., Serra J., “Flat Zones Filtering, Connected Operators, and Filters by Reconstruction”, IEEE Trans. Image Proc., vol. 4, no. 8, Aug 1995

[37] Sarkar S., Majchrzak D., Korimilli K., “Perceptual organization based computational model for robust segmentation of moving objects”, CVIU, no. 86, pp. 141-170, 2002.

[38] Vetro A., Sun H., Wang Y., “MPEG-4 rate control for multiple video objects”, IEEE Trans. Circuits Syst. Video Technol., vol 9, pp. 186-199, Feb. 1999.

[39] Watanabe T, Sasaki Y, Miyauchi S, Putz B, Fujimaki N, Nielsen M, Takino R, Miyakawa S. Attention-regulated activity in human primary visual cortex. *Journal of Neurophysiology*, vol. 79, pp. 2218-2221, 1998.

[40] Yu W., Sommer G., Daniilidis K., “Three dimensional orientation signatures with conic kernel”, Image and Vision Computing, vol. 21, no. 5, pp. 447-458, 2003

[41] “http://homepages.inf.ed.ac.uk/rbf/CAVIAR/”, EC Funded CAVIAR project/IST 2001 37540.

[42] N. Otsu, “A thresholding selection method from gray-scale histogram,” IEEE Trans. on System, Man, and Cybernetics, vol. 9, pp. 62-66, 1979.

[43] J. Serra, “Image Analysis and Mathematical Morphology”, New York: Academic Press, 1982.

[44] Neurmomorphic Vision C++ Toolkit (iNVT), http://ilab.usc.edu/toolkit/, iLab, Univ. Of Southern California.

[45] N. Moenne-Loccoz, E. Bruno, S. Marchand-Maillet, "Knowledge-based Detection of Events in Video Streams from Salient Regions of Activity", Pattern Analysis and Applications, vol. 7, no. 4, pp. 422-429, Dec. 2004

[46] K. Mikolajczyk, C. Schmid, “Scale & Affine Invariant Interest Point Detectors”, International Journal of Computer Vision, vol. 60, no. 1, pp. 63-86, 2004