Spatiotemporal Visual Attention Architecture for Video Analysis

Konstantinos Rapantzikos
School of Electrical & Computer Engineering
National Technical University of Athens
e-mail: rap@image.ntua.gr

Nicolas Tsapatsoulis
Department of Computer Science
University of Cyprus
e-mail: nicolast@ucy.ac.cy

Yannis Avrithis
School of Electrical & Computer Engineering
National Technical University of Athens
e-mail: iavr@image.ntua.gr

Abstract— **Several visual attention (VA) schemes have been proposed with the saliency-based ones being the most popular. The proposed work provides an extension towards VA in video sequences by incorporating the temporal dimension. The architecture is presented in detail and potential applications are investigated. We expect that the extended VA scheme will reveal interesting events across the sequence like occlusions and short occurrences of objects, providing a basis for video surveillance (e.g. intruder detection), segmentation and summarization applications.**

Keywords—visual attention , spatiotemporal processing, scene analysis

I. INTRODUCTION

In most studies, image sequences are processed and analyzed in groups of two frames in order to infer the short-term objects' temporal evolution. Linking together the obtained results generates longer-term dynamics. The actual temporal dimension of the video data is therefore disregarded by incorporating parametric motion model assumptions or smoothing constraints. Spatiotemporal processing may solve several of the problems related to video analysis, [5, 6, 7], but the large amount of data to be processed and the consequent computational burden may keep several researchers from exploiting them. Hence, a mechanism for selecting part of the visual input (*selective visual attention*) to be processed is indispensable for designing successful, computational efficient algorithms in the promising spatiotemporal domain.

The basis of many visual attention models proposed over the last two decades is the feature integration theory of Treisman et al. [1] that was derived from visual search experiments. According to this theory, features are registered early, automatically and in parallel along a number of separable dimensions (e.g. intensity, color, orientation, size, shape etc.). Koch & Ullman [2] have suggested a model based on this theory that leads to the generation of a master saliency map that encodes the saliency of image regions. Meaningful objects (conjunction of features) are identified at a second stage, which requires focused attention. Consequently, at the interface between the first and second stages there should be a bottleneck functioning as a gate allowing only part of the visual information to proceed to the second stage. Itti & Koch proposed the first computational model of saliency-based visual attention [3, 4].

In the current work we propose the extension of the visual attention scheme to volumetric data. Under this framework we treat the video sequence as a video volume with temporal evolution (frame number) being the third dimension. The dimensions of width and height are the usual *x*- and *y*- axes of a frame of video data. The third dimension is derived from layering frames of video data sequentially in time (*x-y-t* space). Consequently, the movement of an object can be regarded as a volume carved out from the 3D space.

II. PROPOSED FRAMEWORK

The extension of Itti *et al.*'s model [3], to the spatiotemporal space described above is presented in this section. The proposed model consists of several intermediate steps and is outlined in Fig. 1. The first processing step consists of cutting the input video into a set of video shots using a common shot-detection technique [11]. After obtaining the spatiotemporal data formation for each shot we provide the extended VA model with input in the form of a video volume. The input volumes are filtered so as to avoid spurious details or noisy areas that will falsely be attended by the proposed system. Feature volumes are then created using a pyramidal decomposition scheme and combined together to generate the final saliency volume.

A. Feature Volume Generation

The main objectives of the first filtering stage are noise removal and simplification of the intensity/color components. We use the *flat-zones* approach [8], to obtain the desired results. The flat-zones are computed by the use of *alternating sequential filters,* which are based on morphological area opening and closing operations with structuring elements of increasing scale [8, 9].

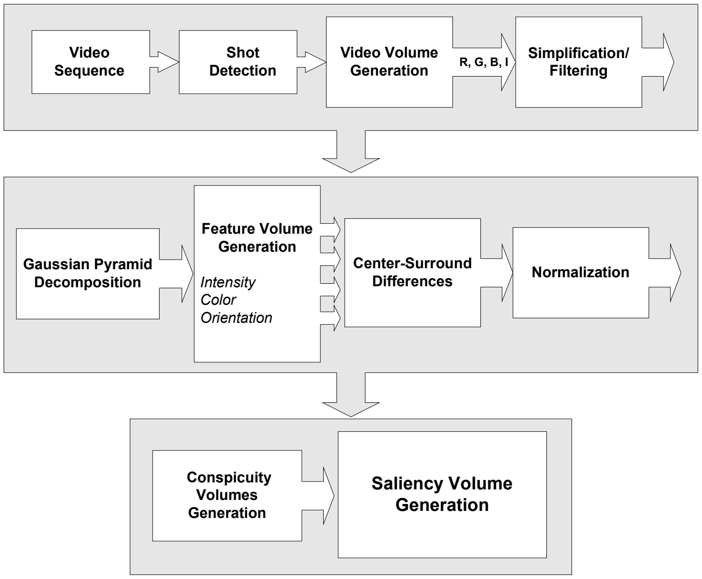
Video Sequence
Shot Detection
Video Volume Generation
R, G, B, I
Simplification/ Filtering
Gaussian Pyramid Decomposition
Feature Volume Generation
Intensity
Color
Orientation
Center-Surround Differences
Normalization
Conspicuity Volumes Generation
Saliency Volume Generation

Figure 1. Overview of the proposed spatiotemporal VA architecture

Each shot is decomposed into four channels: intensity *I*, and color *R, G, B*. Following the computational model of Itti *et al.* we create spatiotemporal dyadic Gaussian pyramids for the intensity *I(σ)* and each color channel *R(σ)*, *G(σ), B(σ)* where *σ* is the spatiotemporal scale. In our implementation, the depth of the pyramid depends on the input video spatiotemporal size, but cannot be less than 6 scales. The actual difference from the 2D case is that each level of the pyramids is actually a subsampled data volume representing the evolution of the observed feature throughout the considered sequence. We generate *feature volumes* for each feature of interest (intensity, color, orientation). Each of them encodes a certain property of the video. Actually, 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.

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, the straightforward extension to spatiotemporal domain is computationally demanding, thus we use a different method in order to obtain similar results in a more computationally efficient way. Hence, we design an algorithm based on morphological tools that shares common ground with the orientation module of the prototype visual attention scheme. Orientation information is obtained from *I* using morphological processing of the corresponding Gaussian pyramids $\mathbf{O}(\theta,\sigma)$, where $\theta \in \{0^o, 45^o, 90^o, 135^o\}$ is the preferred orientation. Accordingly oriented line structuring elements are used in order to obtain the orientations at each frame of the video volume. Superimposing the 2D orientation maps of each separate frame generates the final feature volume.

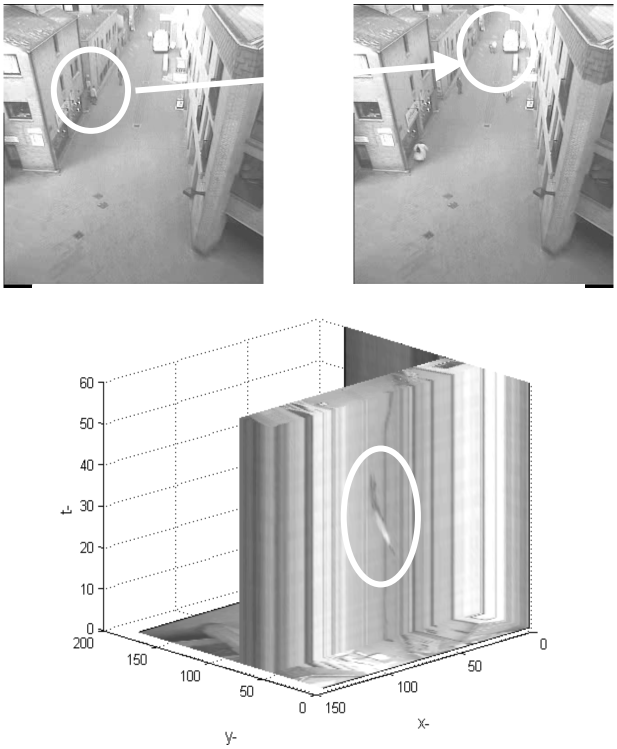

Figure 2. (a) First and last frames of a subsequence showing the trajectory of a pedestrian; (b) corresponding trajectory in 3D space; (regions of interest are encircled)

B. *Saliency Volume Generation*

Biologically inspired "center-surround" differences [3] in three dimensions are used for computing the intensity, color and orientation conspicuity volumes. Center-surround is implemented as the difference between fine and coarse scales of each feature volume. In order to obtain a single volume for each feature we use across-scale differences obtained by interpolation to the finer scale and point-by-point subtraction. Computations are now defined in the 3D space and the point-to-point arithmetic refers to operations between corresponding points in volumes of different spatiotemporal scale. The center is a pixel at scale $c \in \{2,3\}$, and the surround is the corresponding pixel at scale $s = c + \delta$, with $\delta \in \{1,2\}$ (in our implementation).

The first step in combining the different volumes is normalizing them with an operator *N* that promotes the volumes with regions of strong activity in terms of spatiotemporal value. When no knowledge about the scene exists, there is no way to bias the systems towards specific (salient) features. The operator is actually an extended version of the normalization operator presented in [3]. The operator *N* consists of the following: i) normalize all the spatiotemporal feature volumes to the same dynamic range, in order to eliminate across-modality amplitude differences; ii) fr each volume, find the global maximum *M* and the average $\overline{m}$ over all other local maxima; iii) gobally multiply the volume by $(M - \overline{m})^2$. This is an important step since the feature volumes have various dynamic ranges and are obtained by different extraction mechanisms. A single volume for each feature is finally obtained by reduction of each volume to a lower scale and by point-to-point addition. The master saliency volume *S* is generated by simply summing up the obtained conspicuity volumes.

III. Results

In order to visualize and better understand the three-dimensional aspect involved in the proposed architecture we provide a simple illustration that delineates the advantages of VA on the volumetric representation of video. Fig. 2 illustrates a simple scenario of a pedestrian coming out from a building and walking away from the camera (Fig. 2-top). Such a short movement can be readily distinguished from other temporal patterns (non-moving objects) in the video volume as depicted in Fig. 2-bottom. The main idea of visual attention is the extraction of salient regions that pop-out (differ) from the surroundings in terms of specific characteristics. Thus the 3D representation of a video provides a suitable platform for identifying "irregular" (salient) patterns that correspond to meaningful entities. When more than one objects of interest are visible, VA can provide a way to focus the attention on different volumetric shapes, which correspond to regions of decreasing/increasing interest (saliency) in the whole video sequence.

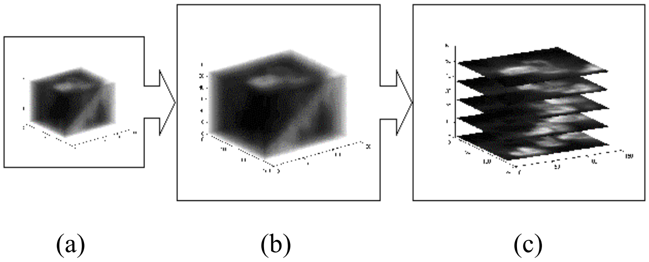

Figure 3. Generation of mask (see text) for visualization purposes. The salient volume is interpolated to the spatiotemporal size of the original sequence and slices for creating the 2D masks

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 semi-transparent mask, which is directly acquired from the corresponding *x-y* slice of the saliency volume. More specifically, the saliency volume of a sequence looks like the one illustrated in Fig. 3a. 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. 3b). Slicing this volume across the temporal dimensions at every time frame produces a saliency map for each of the input frames (Fig. 3c). Superimposing this mask on the corresponding frame generates 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.

The final saliency volume may serve as the preprocessing step for segmenting the video. Regions that pop-out in the spatiotemporal space can be used to enhance the performance of segmentation methods. Considering the limited space, we provide two representative results. The corresponding figures show the original frames and the frames with the superimposed masks in a column-wise order.

The well-known "coast-guard" sequence, depicted in Fig. 4, shows a complex scene with two boats 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 objects' motions are small in magnitude and make the distinction between them rather difficult. A motion segmentation technique should extract the motion vectors and then distinguish by a specific criterion the relative motions of the camera and the boats. This is not simple since the relative motions of two 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. The computationally inefficient motion estimation step

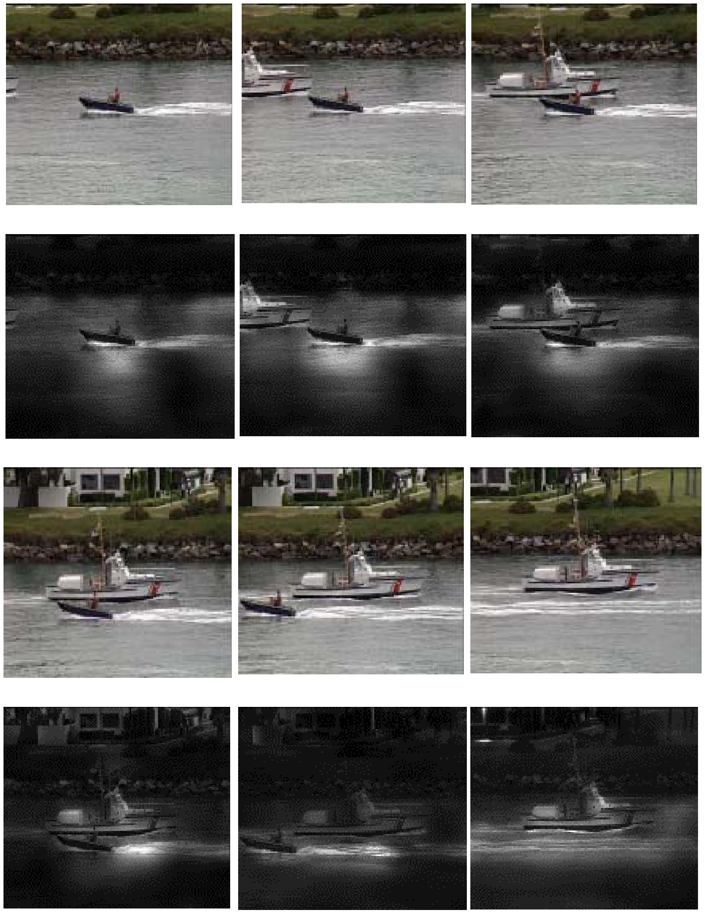

Figure 4. Results on the "coast-guard" sequence. Notice that the global camera motion does not affect the result.

is actually avoided.

Video surveillance systems seek to automatically identify events of interest in a variety of situations. Extracting a salient incoming/outgoing, static/moving object is the most important step of a surveillance system. One specific case is the traffic surveillance one. Generally, the robustness to "noise" conditions affects strongly the existing methodologies. A VA-based system can provide such a system with ROIs in order to process them further and reach important conclusions related to traffic conditions. Knowing that one of the most attractive properties of VA is the robustness to noise, as demonstrated by Itti *et al.* [2], we expect that the proposed system will consistently provide ROIs without being affected by e.g. atmospheric conditions. Results on traffic monitoring videos under varying weather conditions meet our expectations. We show representative results on a sequence acquired by a static camera during a rainy day in Fig. 5.

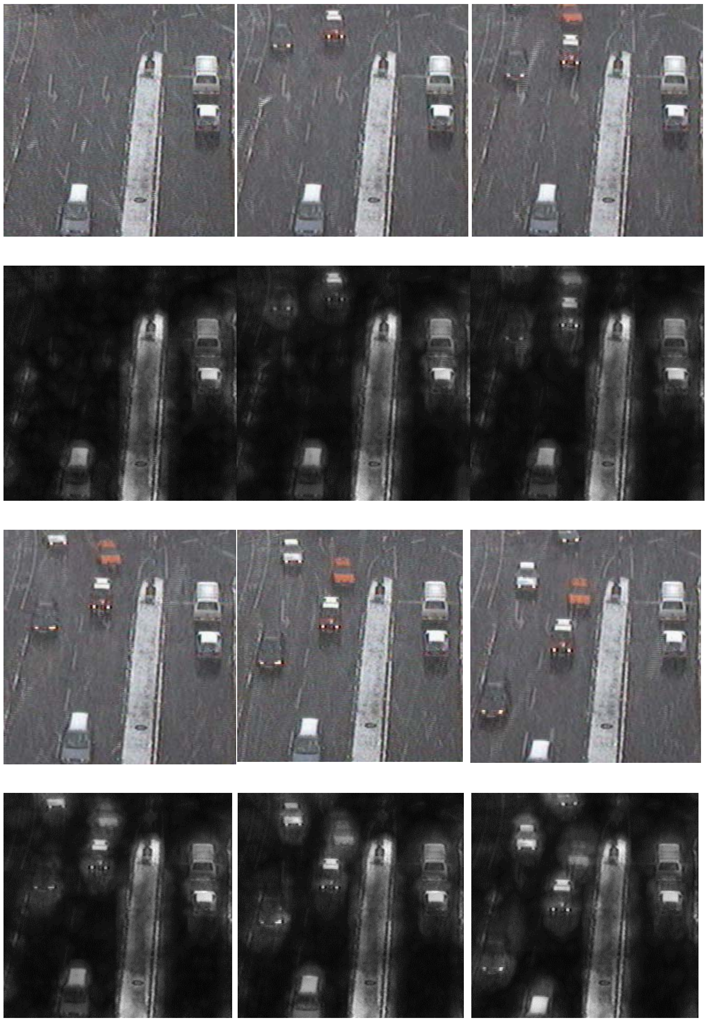

Figure 5. Results on the rainy day traffic surveillance sequence. Notice the incoming car at the left. The corresponding region on the original sequence is of low intensity. Hence, although the car is correctly attended (the mask is transparent) it appears less bright than the others after the mask overlay.

IV. Conclusions

Extracting regions of interest in videos is very important for various applications ranging from video surveillance to retrieval and summarization. VA schemes have proved suitable for static scene processing. We expect that their extension to video in the proposed form will serve as a platform for treating video related processing tasks in a more efficient way.

Based on analysis of the promises of the proposed prototype system, future work will focus on exhaustive experimentation and extensions towards a more robust and informative architecture, in terms of preprocessing results. Such a system could provide e.g. a hierarchical (in terms of saliency) representation of interesting regions in the spatiotemporal space and assist the characterization of video content or the generation of a global index allowing the selective exploration of regions that have different temporal coherence (short/long occurrence), movement (e.g. constant/changing velocity), appearance (specific color & intensity features) etc. Contextual information may also be incorporated in order to enhance the saliency of certain features and trigger scene type-specific top-down reasoning.

References

[1] A. Treisman, "Features and objects in visual processing", *Scientific American* 1986, vol. 255, no. 5, pp. 114-125.

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

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

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

[5] 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.

[6] 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.

[7] 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), Mar 2004.

[8] 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.

[9] 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.

[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] 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.