Handling Uncertainty in Video Analysis with Spatiotemporal Visual Attention

Konstantinos Rapantzikos, Yannis Avrithis, and Stefanos Kollias
School of Electrical & Computer Engineering
National Technical University of Athens, Greece
e-mail: {rap,iavr}@image.ntua.gr, stefanos@cs.ntua.gr

Abstract—**In natural vision, we center our fixation on the most informative points in a scene in order to reduce our overall uncertainty about the scene and help interpret it. Even if we are looking for a specific stimulus around us, we face a great amount of uncertainty since that stimulus could be in any spatial location. Visual attention (VA) schemes have been proposed by researchers to account for the ability of the human eye to quickly fixate on informative regions. Recently, VA in images, and especially saliency-based VA, became an active research topic of the computer vision community. The proposed work provides an extension towards VA in video sequences by integrating spatiotemporal information. The potential applications include video classification, scene understanding, surveillance and segmentation.**

I. Introduction

Despite the common belief that we see everything around us, only a small fraction of the surrounding visual information is processed by the human optical system at any time and aids the reduction of the overall uncertainty about the semantic understanding of the scene. Selecting this small fraction of important information is the main task of the visual attention process (*selective attention*). Processing, analyzing and understanding the visual content of a scene depicted either in a still image or a video sequence presents inherent difficulties, due to camera and object motion, deformations and occlusions. State of the art motion estimation, segmentation, tracking, feature matching, classification and object recognition algorithms are still unable to handle the uncertainty introduced by the above factors in a robust and efficient way. VA can help in this direction by detecting regions of interest and limiting all processing steps in such regions.

Current computational models concern either still images, or video sequences where each input frame is processed sequentially, in a serial manner. Consequently, the regions to be attended are extracted for each frame and the dynamic nature of the attention process can only be inferred by linking together the corresponding results, which is a tedious task. Such methods are prone to noise and can lead to high computational complexity if e.g. motion estimation is one of the prerequisites. It would be desirable to have a general framework that will overcome certain pitfalls and provide a reliable way to analyze the spatial and temporal organization of a video. We believe that an extended VA model, which treats the temporal dimension of a sequence as an intrinsic feature will provide the basis for such a unifying framework. Under such a framework, the reduction in visual uncertainty inferred by locating and analyzing only the interesting events in a sequence will aid the further processing of video data.

Saliency-based selective attention, based on the feature integration theory of Treisman et al. [1], has been computationally modeled in the last decade by Itti and Koch [3, 4], and seems to provide a reasonable first step towards the elucidation and understanding of the visual input. Koch & Ullman [2] have suggested a model based on this theory, leading 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. At the interface between the first and second stages there is a bottleneck functioning as a gate allowing only part of the visual information to proceed to the second stage.

In this paper we elaborate on our previous work, [9], which is based on Itti *et al.'s* scheme. We propose an extension of our spatiotemporal visual attention model by including 3D orientation information and present various interesting results. Under the spatiotemporal framework, we treat the video sequence as a video volume with temporal evolution (frame number) being the third dimension. Specifically, the dimensions of width and height are the usual *x*- and *y*- axes of a video frame. The third dimension (depth) is derived by layering the frames sequentially in time, constructing a *x-y-t* space. Consequently, the movement of a region or object can be regarded as a volume carved out from the 3D space. It has to be mentioned that the proposed model is limited to bottom-up control of attention, like the one of Itti *et al.*, which means that no volitional component (e.g. *a priori* knowledge) is incorporated. Furthermore, we are only concerned with the localization of the events to be attended and not their identification.

Section 2 of this paper describes the architecture of the proposed spatiotemporal VA framework, as well as its extension with 3D orientation features. Section 3 provides a set of experimental results on video sequences to illustrate the performance of the proposed model, comparing to other approaches. Finally, conclusions are drawn in Section 4.

II. SPATIOTEMPORAL VA FRAMEWORK

The architectrure of the proposed spatiotemporal VA model is presented below, including all intermediate processing steps. The extension of this model with the addition of 3D orientation information to handle motion in the spatiotemporal domain is discussed next.

A. VA Architecture

Given an arbitrary input sequence, the first processing step consists of slicing it into a set of video shots using a common shot-detection technique [8]. 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 that is sufficient to adequately represent object trajectories. The acquired frames form a video volume as explained above. This volume is decomposed into a set of distinct "channels" by using linear filters tuned to specific stimulus dimensions, such as luminance, red, green, blue, yellow hues and various orientations. 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, as explained in [2]. Each of these *feature volumes* encodes a certain property of the video.

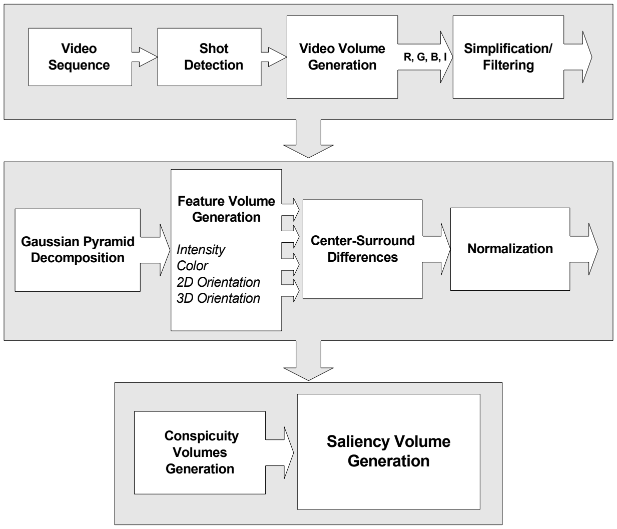
Video Sequence
Shot Detection
Video Volume Generation
R, G, B, I
Simplification/ Filtering
Gaussian Pyramid Decomposition
Feature Volume Generation
Intensity
Color
2D Orientation
3D Orientation
Center-Surround Differences
Normalization
Conspicuity Volumes Generation
Saliency Volume Generation

Fig. 1. Spatiotemporal VA: step-by-step.

After obtaining the spatiotemporal data formation, the input volumes are morphologically filtered by a flat zone approach as in [6], so as to avoid spurious details or noisy areas that might otherwise be erroneously attended by the proposed system. Afterwards, following the structure of the static image-based approach of Itti & Koch, we generate feature volumes for each feature of interest, including intensity, color and 2D/3D orientation, as explained in section 2.1. Each of them encodes a certain property of the video. Every volume simultaneously represents the spatial distribution and temporal evolution of the encoded feature. A *normalization* operator is responsible for enhancing the most salient subvolumes inside them so as to prohibit non-important regions from drastically affecting the result.

The process described above is performed at a number of different spatiotemporal scales, to allow the model to represent smaller and larger "events" in separate subdivisions of theses channels. This multiple scale representation is obtained through Gaussian pyramids. Center-surround operations, which are suitable for detecting locations that locally stand out from their surroundings, are implemented as differences between a fine and a coarse scale for a given feature. Finally, a linking stage fuses the separate volumes and produces a salient one that represents interesting events as enhanced (in terms of intensity) spatiotemporal regions. Fig. 1 illustrates all intermediate steps of the proposed model.

B. 3D Orientation

Orientation in a spatiotemporal domain can be calculated in 2D by extracting oriented edges at each frame and superimposing the results, as indicated in [9], and by direct 3D filtering of the video volume. 3D filtering is related to motion analysis tasks since orientation in space-time corresponds to velocity [7]. In order to get the 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, [11], it is not included as a feature map in the saliency-based computational model of Itti *et al.* [3]. Elsewhere we have used motion for event analysis purposes [10]. Actually, motion/velocity description of the objects can be directly extracted by the 3D orientation volume as described in [12], therefore avoiding the need for independent optical flow computation.

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 to achieve computationally efficient results. We filter the volume with rotated versions of an orientation-selective morphological structuring element and produce a

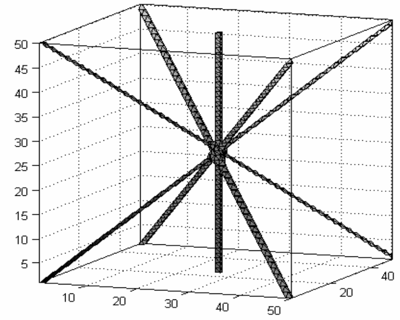

Fig. 2. Five of the nine main 3D orientations used; not all of them are shown for illustrative purposes.

result with enhanced oriented subvolumes being the result of the objects' path in the scene. 3D Cylinder-shaped structuring elements are used in order to obtain the desired 3D orientations of the video volume. Five of the nine main orientations used are illustrated in Fig. 2.

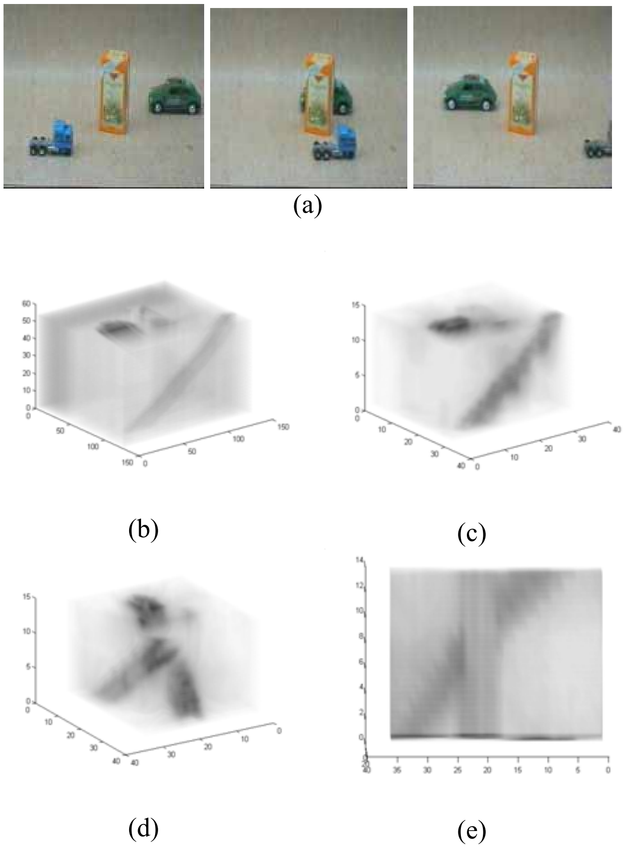

Fig. 3 (a) Three representative frames of the sequence; (b) unprocessed video volume; (c)-(e) saliency volume observed from 3 different angles. The volumes are negative and transparent versions of the original saliency volume for visualization purposes.

III. RESULTS

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. Three representative frames of the sequence are shown in Fig. 3a, the semi-transparent volume of the original sequence in Fig. 3b, and the saliency volume under three different angles Fig. 3c-e. 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.

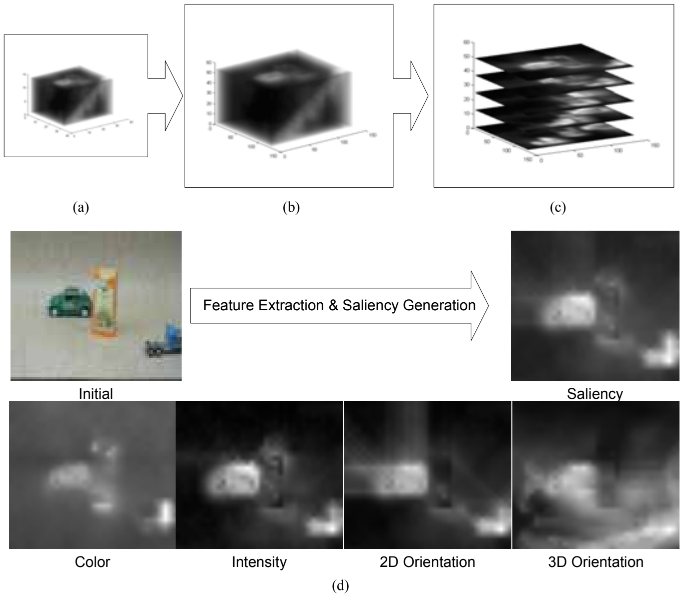

Fig. 4. (a)-(c) Generation of mask (see text) for visualization purposes; (d) the initial frame and corresponding slices for each of the feature and saliency volumes.

Despite of the simplicity of the previous example, it becomes obvious that the uncertainty involved in the understanding of the current scene can be greatly reduced. A segmentation or recognition algorithm should process only the salient regions (subvolumes) of the initial video.

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. 4a. 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. 4b). Slicing this volume across the temporal dimensions at every time frame produces a saliency map for each of the input frames (Fig. 4c). Superimposing this map on the corresponding frame generates the desired result, as shown in Fig. 4ds. 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 *"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. 5 show the original frame and the corresponding saliency mask derived from the saliency volume, respectively. 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 an object's position or for improving an incremental tracking approach [13, 14]. 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 [13]. The magnitude of the motion field (square root of motion vectors in x- and y-directions) generated by Black & Anandan's method [15] is shown in the third column of Fig. 5. 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.

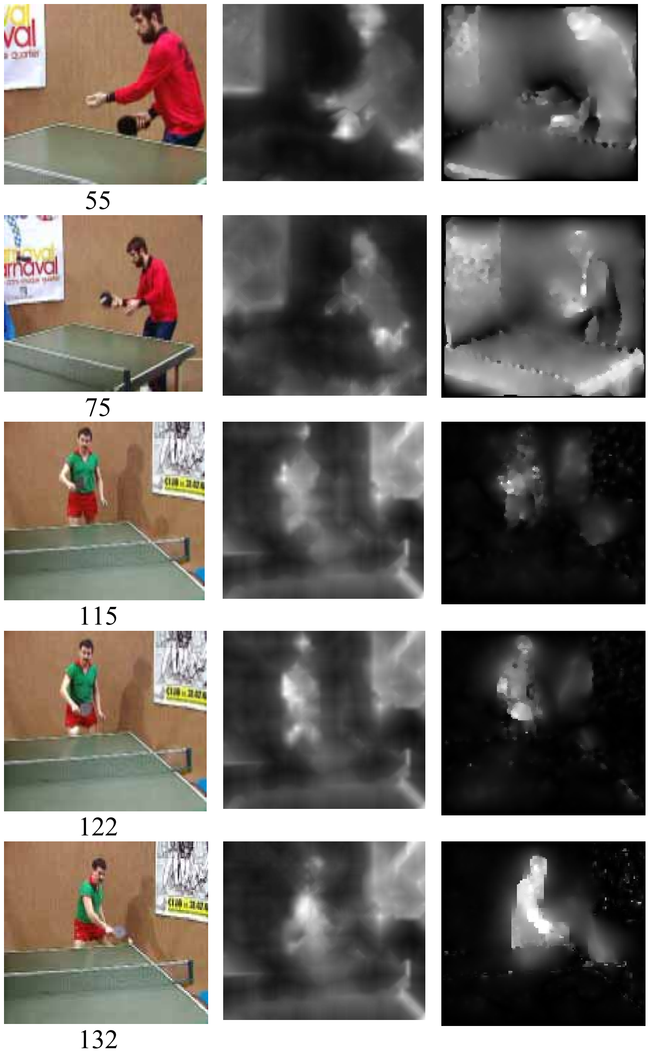

Fig. 5. Results on the "table-tennis" sequence (numbers correspond to frames). Row-wise: original frame, saliency map and magnitude of the motion map

IV. CONCLUSIONS

Extracting regions of interest in video is very important for various applications ranging from video surveillance to retrieval and summarization. VA schemes have been proved to be suitable for static scene processing. We expect that their extension to video, as proposed in [9], with the addition of the 3D orientation information will serve as a platform for treating video related processing tasks in a more efficient way.

Further experimentation is required to further prove the efficiency of the implemented model and put up new applications in the field including segmentation, tracking and summarization. Nevertheless, the proposed model is limited to bottom-up control of attention. Furthermore, we are only concerned with the localization of the events to be attended and not their identification. Future work should focus on the incorporation of a top-down component (*a priori* knowledge) in order to select regions not only due to their saliency but also by means of semantics related to the scene.

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] 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.

[6] 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.

[7] 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.

[8] 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.

[9] Rapantzikos K., Tsapatsoulis N., Avrithis Y., "Spatiotemporal Visual Attention Architecture for Video Analysis Proc. of IEEE International Workshop On Multimedia Signal Processing (MMSP'04), Sienna, 2004

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

[11] 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.

[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] 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.

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

[15] 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