On the implementation of Visual Attention Architectures

KONSTANTINOS RAPANTZIKOS AND NICOLAS TSAPATSOULIS

DEPARTMENT OF ELECTRICAL AND COMPUTER ENGINEERING

NATIONAL TECHNICAL UNIVERSITY OF ATHENS

9, IROON POLYTECHNIOU STR., 15773 ZOGRAFOU, GREECE

Abstract:

The majority of visual attention models is based on the concept of *saliency map*, a two-dimensional map that encodes the saliency of objects in the surrounding world. We attempt a short review of current implementations and present our first thoughts on extending the classical saliency-based model by using motion and prior knowledge.

Keywords: visual attention, saliency map, attention control

Introduction

Most of our impressions and memories are based on the sense of vision. However, the biological mechanisms involved in the human vision are not obvious either for the observer or the experienced researcher. How do we understand shape and motion? How do we perceive color? The problem becomes more complex due to the fact that the human brain receives information for shape, motion and color through (at least) three different, parallel and interrelated processing routes [1]. The latter fact arises the second equally complex problem of combining these three inputs to form a single image of the world. In an attempt to understand visual perception numerous experiments have been performed in order to discover brain regions involved in various aspects of vision. Although, the linking of the available information is still a mystery, Ann Treisman [2] proved that the formation of the possible links requires *attention*. Operationally, information can be said to be "attended" if it enters short-term memory and remains there long enough to be voluntarily reported. Thus, visual attention is closely linked to *visual awareness* [4]. The brain attentional mechanism involved in analyzing and

recognizing the surroundings distinguishes objects by focusing on elementary properties: brightness, color and orientation.

Attention control has been found to arise by two mechanisms [2, 3]: a bottom-up one that biases the observer towards selecting stimuli based on their *saliency*, and a top-down one that directs the "spotlight of attention" under cognitive, volitional control. The control network deciding between the importance of desired (top-down) and unexpected (bottom-up) sites for attention is still unexplored.

Koch and Ullman, [9], introduced the idea of a saliency map to accomplish preattentive selection. This is an explicit two-dimensional map that encodes the saliency of objects in the visual environment. Competition among neurons in this map gives rise to a single winning location that corresponds to the most salient object, which constitutes the next target. If this location is subsequently inhibited, the system automatically shifts to the next most salient location, endowing the search process with internal dynamics.

The investigation presented in this paper aims at a short description of the saliency-based visual attention and at the integration of additional information into it towards volitional control (top-down).

Saliency-based model of visual attention

The original version of the saliency-based model of visual attention presented in [4] deals with static color images. Visual input is first decomposed into a set of topographic feature maps. Different spatial locations then compete for saliency within each map, such that only locations which locally stand out from their surround can persist. All feature maps feed, in a purely bottom-up manner, into a master saliency map, which topographically codes for local conspicuity over the entire visual scene. In primates, such a map is believed to be located in the posterior parietal cortex [6] as well as in the various visual maps in the pulvinar nuclei of the thalamus [7].

Itti and Koch [4, 8] presented an implementation of the proposed saliency-based model. Low-level vision features (color channels tuned to red, green, blue and yellow hues, orientation and brightness) are extracted from the original color image at several spatial scales, using linear filtering. The different spatial scales are created using Gaussian pyramids, which consist of progressively low-pass filtering and sub-sampling the input

image. Each feature is computed in a center-surround structure akin to visual receptive fields. Using this biological paradigm renders the system sensitive to local spatial contrast rather than to amplitude in that feature map. Center-surround operations are implemented in the model as differences between a fine and a coarse scale for a given feature. Seven types of features, for which evidence exists in mammalian visual systems, are computed in this manner from the low-level pyramids. The implemented algorithm is summarized in Figure 1 (central part).

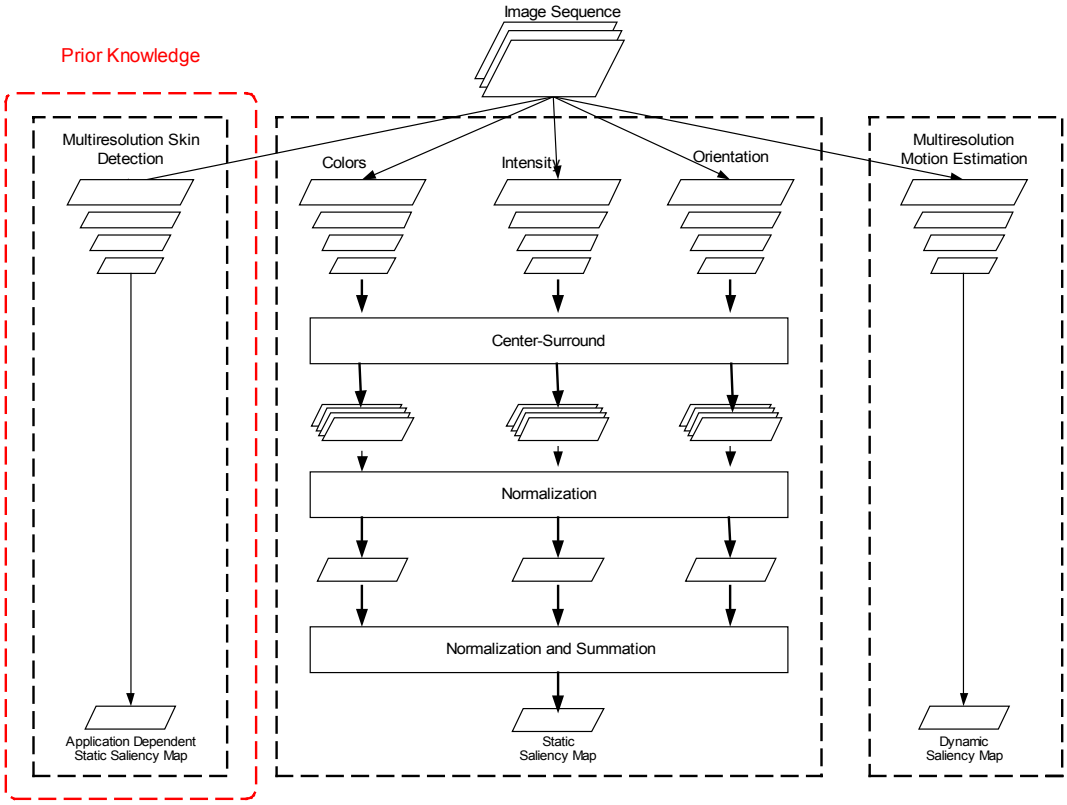

Figure 1: Schematic diagram of the saliency-based visual attention including motion and skin detection module

Integration of Motion into the Model

Motion is of fundamental importance in biological vision systems and contributes to visual attention as confirmed by Watanabe et al. in [10]. Despite the biological evidences, only few researchers studied the integration of motion into the saliency-based model. Several authors [11, 12, 13] attempted the integration of dynamic features, but their methods lack the interaction between the two different feature classes (static-dynamic) to build a global attention map.

We use a multiresolution gradient-based approach, [14], to estimate optical flow and generate a new conspicuity map in the same manner as with static maps (Figure 1-right part). Although not fully covered due to lack of space, we think of two possible ways to combine dynamic and static features: Either by a weighted summation of the available maps or by prioritizing the motion information and attending only the moving objects.

Integration of top-down information into the model

It had been thought that bottom-up signals normally achieved attention capture; it is now appreciated that top-down control is usually in charge. Involuntary attention capture by distracting inputs occurs only if they have a property that a person is using to find a target [15]. Towards this direction we attempt to integrate prior knowledge to the saliency-based model in order to draw the attention to regions with specific characteristics (Figure 1-left part). As an example we consider the face detection case. We use a skin detector scheme to generate a skin map and link it with the other feature maps.

Experimental results

The extended saliency-based model was tested with a real sequence showing a man moving his arms and hands in front of a static background. Figure 2 shows the generated feature maps and the different image characteristics that each of them captures. Obviously, the motion map provides important information by distinguishing between moving and non-moving objects. Additionally, the skin map exhibits high activation (bright areas) at regions with the desired property. The last row of Figure 2 shows the saliency maps computed using the classical approach and the extended one. In Figure 2(h) the objects of interest, namely the face and the moving arms/hands stand out and the salience map is less affected by non-uniform illumination and reflections observed in Figure 2(g).

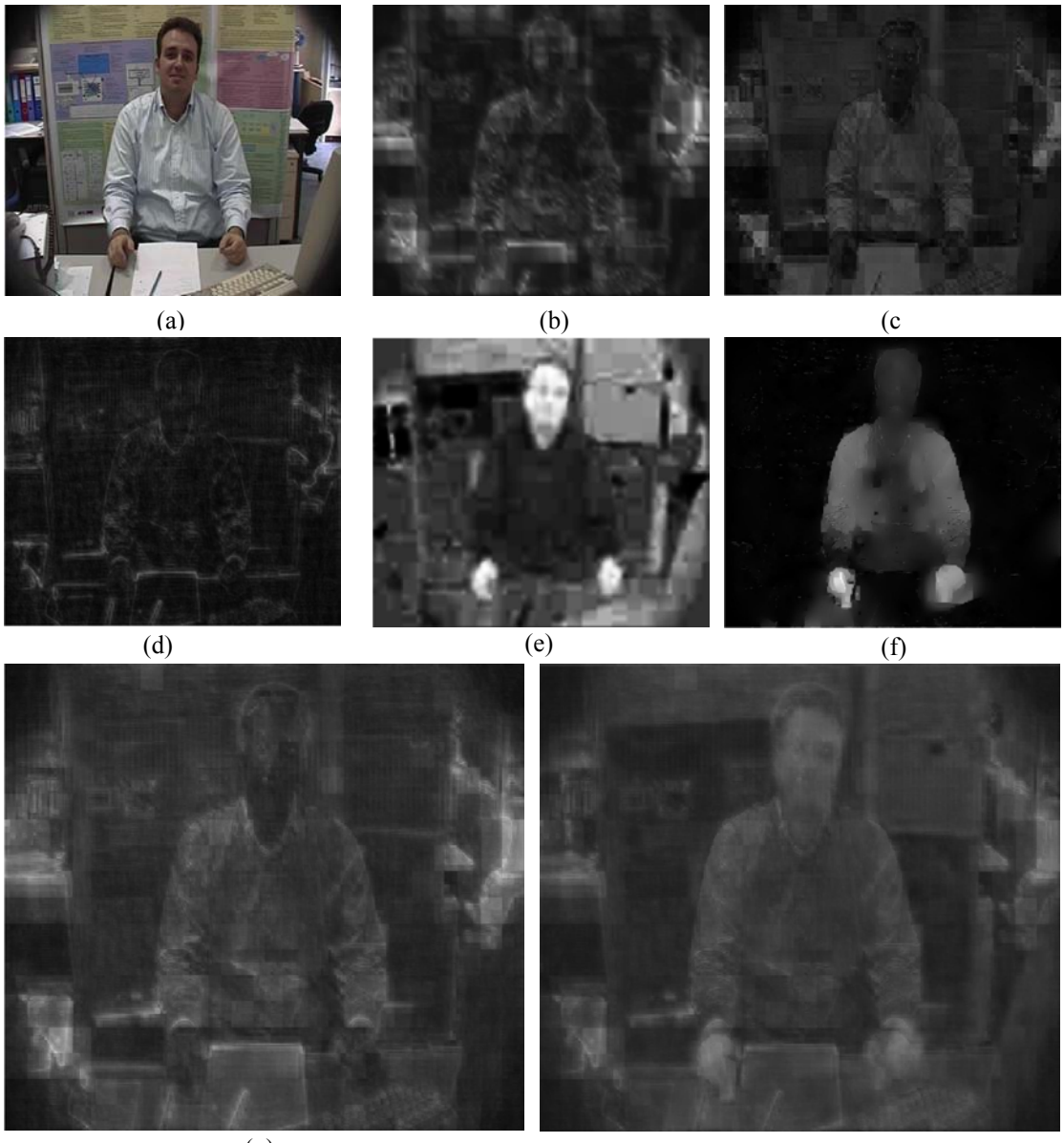

Figure 2: (a) original image (in color); (b) intensity map; (c) color map; (d) orientation map; (e) skin map; (f) motion map; (g) intensity/color/orientation saliency map; (h) intensity/color/ orientation/ motion/skin saliency map

References:

1. Kandel E.R., Schwartz H.J., Jessell M.T. Essentials of Neural Science and Behavior. *Appleton & Lange*, 1995

2. Treisman A. Features and objects in visual processing. *Scientific American*, 255(5): 114-125, 1986.

3. Bergen J.R., Julesz B., Parallel versus serial processing in rapid pattern discrimination. *Nature*, 303, pp. 696-698, 1983.

4. 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.

5. Crick F., Koch C.. Constraints on cortical and thalamic projections: the no-strong-loops hypothesis. *Nature*, 391, pp. 245-250, 1998.

6. Gottlieb J.P., Kusunoki M., Goldberg M.E.. The representation of visual salience in monkey parietal cortex. *Nature*, vol. 391, no. 6666, pp. 481-484, 1988.

7. Robinson D.L., Peterson S.E. The pulvinar and visual salience. *Trends in Neuroscience*, vol. 15, no. 4, pp. 127-132, 1992.

8. 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.

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

10. 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.

11. Tsotsos J.K., Culhane S.M., Wai W.Y.K., Lai Y.H., Davis N., Nuflo F.. Modelling visual attention via selective tuning. *Artificial Intelligence*, vol. 78 (1-2), pp. 507-545, 1995.

12. Maki A., Nordlund P., Eklundh J.O. Attentional scene segmentation: Integrating depth and motion from phase. *Computer Vision and Image Understanding*, vol. 78, pp. 351-373, 2000.

13. Milanese R., Gil S., Pun T. Attentive mechanisms for dynamic and static scene analysis. *Optical Engineering*, vol. 34, no. 8, pp. 2428-2434, 1995.

14. Black M.J., Anandan P. The Robust Estimation of Multiple Motions: Parametric and Piecewise-Smooth Flow Fields. *CVIU*, 63(1), pp. 75-104, 1996.

15. Pashler H. Attention and performance. *Ann. Rev. Psych*, vol. 52, pp. 629-651, 2001