Available online at www.sciencedirect.com
SCIENCE DIRECT®
Vision Research 45 (2005) 205–231
Vision Research
www.elsevier.com/locate/visres

Modeling the influence of task on attention

Vidhya Navalpakkam, Laurent Itti *

Departments of Computer Science, Psychology and Neuroscience Graduate Program, University of Southern California, Hedco Neuroscience Building, Room 30A, Mail Code 2520, 3641 Watt Way, Los Angeles, CA 90089-2520, USA

Received 6 October 2003; received in revised form 19 March 2004

Abstract

We propose a computational model for the task-specific guidance of visual attention in real-world scenes. Our model emphasizes four aspects that are important in biological vision: determining task-relevance of an entity, biasing attention for the low-level visual features of desired targets, recognizing these targets using the same low-level features, and incrementally building a visual map of task-relevance at every scene location. Given a task definition in the form of keywords, the model first determines and stores the task-relevant entities in working memory, using prior knowledge stored in long-term memory. It attempts to detect the most relevant entity by biasing its visual attention system with the entity's learned low-level features. It attends to the most salient location in the scene, and attempts to recognize the attended object through hierarchical matching against object representations stored in long-term memory. It updates its working memory with the task-relevance of the recognized entity and updates a topographic task-relevance map with the location and relevance of the recognized entity. The model is tested on three types of tasks: single-target detection in 343 natural and synthetic images, where biasing for the target accelerates target detection over twofold on average; sequential multiple-target detection in 28 natural images, where biasing, recognition, working memory and long term memory contribute to rapidly finding all targets; and learning a map of likely locations of cars from a video clip filmed while driving on a highway. The model's performance on search for single features and feature conjunctions is consistent with existing psychophysical data. These results of our biologically-motivated architecture suggest that the model may provide a reasonable approximation to many brain processes involved in complex task-driven visual behaviors.

Keywords: Attention; Top-down; Bottom-up; Object detection; Recognition; Task-relevance; Scene analysis

1. Introduction

There is an interesting diversity in the range of hypothetical internal scene representations, including the *world as an outside memory* hypothesis that claims no photographic memory for visual information (O'Regan, 1992), the *coherence theory* according to which only one spatio-temporal structure or coherent object can be represented at a time (Rensink, 2000), a limited memory of three or four objects in visual short-term memory (Irwin & Andrews, 1996; Irwin & Zelinsky, 2002), and finally, memory for many more previously attended objects in visual short-term and long-term memory (Hollingworth, 2004; Hollingworth & Henderson, 2002; Hollingworth, Williams, & Henderson, 2001). Together with studies in change detection (Kanwisher, 1987; Rensink, 2000, 2002; Rensink, O'Regan, & Clark, 1997; Watanabe, 2003), this suggests that internal scene representations do not contain complete knowledge of the scene. To summarize, instead of attempting to segment, identify, represent and maintain detailed memory of all objects in a scene, there is mounting evidence that our brain

* Corresponding author. Tel.: +1 213 740 3527; fax: +1 213 740 5687.
E-mail addresses: navalpak@usc.edu (V. Navalpakkam), itti@usc.edu (L. Itti).

0042-6989/$ - see front matter © 2004 Elsevier Ltd. All rights reserved.
doi:10.1016/j.visres.2004.07.042

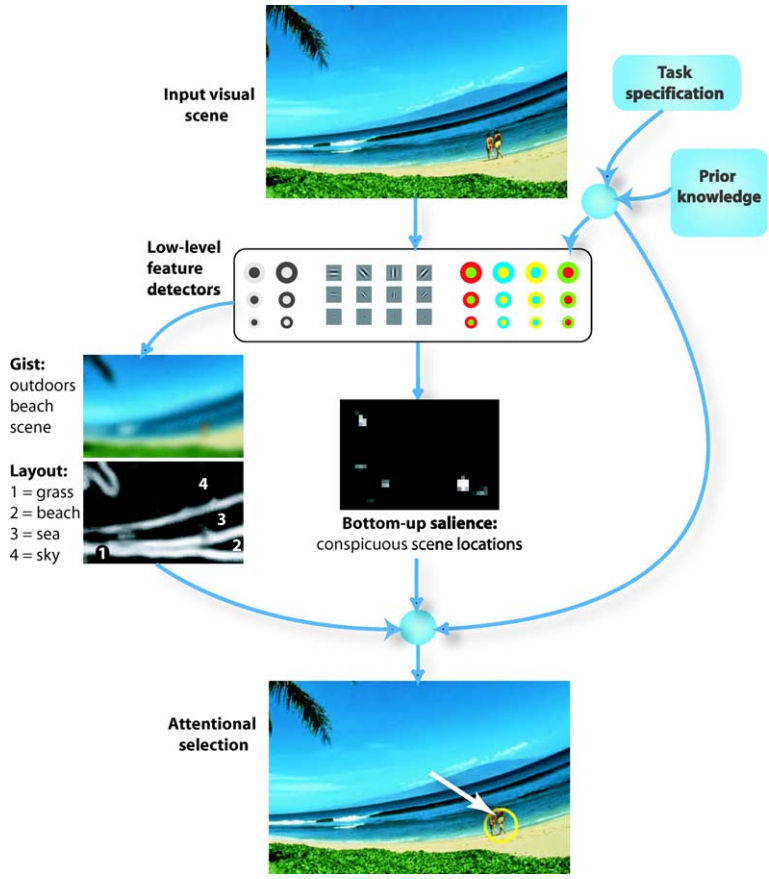

Fig. 1. Overview of current understanding of how task influences visual attention: Given a task such as "find humans in the scene", prior knowledge of the target's features is known to influence low-level feature extraction by priming the desired features. These low-level features are used to compute the gist and layout of the scene as well as the bottom-up salience of scene locations. Finally, the gist, layout and bottom-up salience map are somehow combined with the task and prior knowledge to guide attention to likely target locations. The present study attempts to cast this fairly vague overview model into a more precise computational framework that can be tested against real visual inputs.

may adopt a *need-based* approach (Triesch, Ballard, Hayhoe, & Sullivan, 2003), where only desired objects are quickly detected in the scene, identified and represented.

How do we determine the desired objects, and isolate them from within around 10^8 bits of information bombarding our retina each second? In this section, we provide a brief overview of some crucial factors. A detailed review of relevant literature can be found in Section 2. Studies of eye movements, physiology and psychophysics show that several factors such as bottom-up cues, knowledge of task, gist of the scene, [1] and nature of the target play important roles in selecting the focus of attention (see Fig. 1 for current understanding). Bottom-up processing guides attention based on image-based low-level cues. Such processes make a red ball more salient among a set of black balls. Gist and layout [2] guide attention to likely target locations in a top-down manner, e.g., if the task is to find humans in the scene and the gist is an outdoor beach scene, humans can be found by focusing attention near the water and the sand. Prior knowledge of the target also accelerates target detection in visual search tasks and this suggests that our visual system biases the attentional system with the known target representation so as to make the target more salient. Further, the classic eye movement experiments of Yarbus (1967) show drastically different patterns of eye movements over a same scene, depending on task. To summarize, task (with the aid of the gist and knowledge of the target) plays an important role in the selection of the focus of attention. As a consequence, eye movements vary depending on the task

[1] An abstract meaning of the scene that refers to semantic scene category, such as indoor office scene, outdoor beach scene etc.

[2] Division of the scene into regions in space based on semantic or visual similarity, e.g., a typical beach scene consists of three regions—sky on top, water in the middle, and sand at the bottom.