Available online at www.sciencedirect.com
ScienceDirect
BioSystems 86 (2006) 91–99
Bio Systems
www.elsevier.com/locate/biosystems

Modeling feature-based attention as an active top-down inference process

Fred H. Hamker*

Allgemeine Psychologie, Psychologisches Institut II, Westf. Wilhelms-Universität, Fliednerstrasse 21, 48149 Münster, Germany

Received 15 December 2005; received in revised form 14 March 2006; accepted 17 March 2006

Abstract

Vision is a crucial sensor. It provides a very rich collection of information about our environment. The difficulty in vision arises, since this information is not obvious in the image, it has to be constructed. Wheres earlier approaches have favored a bottom-up approach, which maps the image onto an internal representation of the world, more recent approaches search for alternatives and develop frameworks which make use of top-down connections. In these approaches vision is inherently a constructive process which makes use of a priory information. Following this line of research, a model of primate object perception is presented and used to simulate an object detection task in natural scenes. The model predicts that early responses in extrastriate visual areas are modulated by the visual goal.

Keywords: Vision; Top-down; Bottom-up; Feature-based attention; Computational model; Gain control; V4

1. Introduction

Object recognition, generally implemented in a hierarchical bottom-up process (Fukushima, 1980; Perrett and Oram, 1993; Wallis and Rolls, 1997; Riesenhuber and Poggio, 1999) in which the complexity of representation along with the receptive field size increases, leads to a strong overlapping of populations encoding features belonging to different objects. These ambiguities in cell populations encoding features within the same receptive field limit the use of these approaches for non-segmented scenes like natural images.

The closely linked paradigms of active vision, purposive vision and animate vision (Aloimonos, 1993; Ballard, 1991) have proposed that bottom-up directed vision is an ill-posed-problem and suggested each task requires its own specific algorithm. In this regard, an universal, general vision is not possible. According to these paradigms, the fundamental problem of vision is the selection of the relevant information within the scene and the computation of an appropriate representation. An "active" vision system – in the sense of a visually selective device – is able to acquire the necessary information on demand by focusing on the relevant areas within the visual scene and taking different views from the same object.

The approach of "Deictic Codes for the Embodiment of Cognition" aims to provide a framework for describing the phenomena that appear at about one-third of a second in the perception–action process (Ballard et al., 1997). Deictic primitives dynamically refer to points in the world with respect to their crucial describing features (e.g., color or shape). The outcome of the processing after one-third second, which is the natural sequentiality of body movements can be matched to the natural computational economies of sequential decision systems through a system of implicit reference (called deictic) in which pointing movements are used to bind objects in the world

* Tel.: +49 251 8334171; fax: +49 251 8334180.
E-mail address: fhamker@uni-muenster.de (F.H. Hamker).

0303-2647/$ – see front matter © 2006 Elsevier Ireland Ltd. All rights reserved.
doi:10.1016/j.biosystems.2006.03.010

to cognitive programs. Ballard et al. (1997) suggested visual routines (Kosslyn, 1994; Ullman, 1984; Just and Carpenter, 1976) to divide one complex task into subtasks, such as selection and identification.

Selective perception has been addressed in attention related experimental frameworks such as visual search. The basic idea is that once an object is selected by a focus of attention it can be connected to an internal pointer and being processed in high-level areas. This view has its origin in the classical approach of perception that separates between a pre-attentive and attentive stage (Treisman and Gelade, 1980). Computer implementations of these types of models use a saliency map to indicate a location of interest (Koch and Ullman, 1985; Wolfe, 1994; Itti and Koch, 2000) and compute a focus of attention that selects an object (Olshausen et al., 1993). This focus could be guided by some rough knowledge about an object, such as its color. Feature-based attention is left to only guide the selection process by weighting the input into the saliency map (Wolfe, 1994; Milanese et al., 1995; Navalpakkam and Itti, 2005).

We have developed an alternative approach in which feature-based attention acts on the object representations itself. Spatially selective attentive binding, however, occurs through reentrant oculomotor loops. The search for an object or just parts of it produces top-down expectations, which meet the bottom-up processed stimulus features in the ventral pathway. This initiates a dynamic and distributed recognition process at different levels of the hierarchy by enhancing the features of interest. At higher areas these are typically complex patterns. At lower levels these complex patterns have to be decomposed into more simple patterns. Thus, top-down inference has to rely on reverse weights to decompose a pattern into its parts. By competitive interactions such a mechanism would allow to flexibly filter out the information which is inconsistent with the high-level goal description. However, the sensory evidence of the encoded items does not always allow to rule out all objects but one. This top-down inference only strengthens the expected features, which are not necessarily the to be reported ones, and guides goal-directed behavior. Thus, in parallel, areas responsible for oculomotor selection start to plan appropriate responses. Specifically, the target location of the planed eye movement is used for a location specific inference operation which in turn filters out objects at irrelevant locations. This spatial attention effect could be interpreted as a shortcut of the actual planned eye movement. It facilitates planning processes to evaluate the consequences of the planned action. As a result of both inference operations, the high-level goal description is bound to an object in the visual world.

In this approach vision is an active, dynamic and constructive process. It allows a more close look onto the processes of binding objects in the world to cognitive programs that act within one-third of a second. Our proposed concept relies on top-down connections in vision, which have been discussed and its usefulness has been demonstrated for several times (Grossberg, 1980; Mumford, 1992; Ullman, 1995; Tononi et al., 1992; Tsotsos et al., 1995; Rao and Ballard, 1999; Rao, 1999; Hamker, 1999; Engel et al., 2001; Hamker and Worcester, 2002; Corchs and Deco, 2002; Hochstein and Ahissar, 2002; Rao, 2004; Hamker, 2004b). However, top-down connections have not been used in an unequivocal fashion. The generative approach (Mumford, 1992; Olshausen and Field, 1997; Rao, 1999) predicts that the top-down signal is subtracted from the bottom-up signal. Such models predict a reduction of activity when the predicted input matches with the actual input. Our model predicts an enhancement, as previously suggested by ART (Grossberg, 1980). We have shown that this is consistent with cell recordings in IT, V4 and FEF in visual search (Hamker, 2005a) and in other attentional experiments (Hamker, 2004a,b). Since these simulations have been done with artificial inputs, we have recently scaled up this model to simulate object detection (Hamker, 2005c,d) and change detection (Hamker, 2005b) tasks in natural scenes. Here, we will focus on feature-based attention in area V4/TEO with respect to the search task.

2. The model

2.1. Anatomical, pysiological and behavioral evidence

The brain has developed specific functional areas in the visual cortex, which can be divided into two major streams. Form and color travel from V1 to V2, V4 of the occipital lobe into TEO and TE of the inferior temporal lobe (Zeki, 1978; Livingstone and Hubel, 1988). This ventral pathway is known to encode object identity. It is generally accepted that the complexity of encoded features increases along the ventral pathway. V1 neurons can be driven by simple properties of a stimulus, such as the orientation of a bar. TE neurons, however, encode highly sophisticated shape properties. These "experts" have probably evolved to meet the statistics of stimuli we typically encounter. The receptive field size has also been suggested to increase along the ventral pathway as well. Most of the receptive field size mappings have been done with anestesized monkeys. The idea is that the increasing receptive field size supports location