Robotics and Autonomous Systems 71 (2015) 35–48

Contents lists available at ScienceDirect

Robotics and Autonomous Systems

journal homepage: www.elsevier.com/locate/robot

A developmental where–what network for concurrent and interactive visual attention and recognition

Zhengping Ji [a,*], Juyang Weng [b]

[a] Advanced Image Research Lab (ARIL), Samsung Semiconductor Inc., Pasadena, CA, 91103, United States

[b] Department of Computer Science and Engineering, Michigan State University, East Lansing, MI 48824, United States

HIGHLIGHTS

- Computational modeling of attention and recognition as ubiquitous internal actions.
- Development of sensory invariance through motor-specific entropy reduction.
- Spatiotemporal optimization of cortical self-organization.
- Non-iterative bidirectional sparse coding model.
- Divide-and-conquer solution to learn deep neural network with bidirectional flows.

ARTICLE INFO

Article history:
Available online 27 March 2015

Keywords:
Developmental learning
Where–what sensorimotor pathways
Attention
Recognition
Brain-inspired neural network

ABSTRACT

This paper presents a brain-inspired developmental architecture called Where–What Network (WWN). In this second version of WWN, WWN-2 is learned for concurrent and interactive visual attention and recognition, via complementary pathways guided by "type" motor and "location" motor. The motor-driven top-down signals, together with bottom-up excitatory activities from the visual input, shape three possible information flows through a Y-shaped network. Using ℓ_0 constrained sparse coding scheme, the top-down and bottom-up co-firing leads to a non-iterative cell-centered synaptic update model, entailing the strict entropy reduction from early to later layers, as well as a dual optimization of update directions and step sizes that dynamically depend on the firing ages of the neurons. Three operational modes for cluttered scenes emerge from the learning process, depending on what is available in the motor area: context-free mode for detection and recognition from a cluttered scene for a learned object, location-context mode for doing object recognition, and type-context mode for doing object search, all by a single network. To demonstrate the attention capabilities along with their interaction of visual processing, the proposed network is in the presence of complex backgrounds, learns on the fly, and produces engineering graded performance regarding attended pixel errors and recognition accuracy. As the proposed architecture is developmental, meaning that the internal representations are learned from pairs of input and motor signal, and thereby not manipulated internally for a specific task, we argue that the same learning principles and computational architecture can be potentially applicable to other sensory modalities, such as audition and touch.

1. Biological visual pathway and functions

Studies in neuroscience have identified two main pathways in the primate vision system, the ventral pathway and the dorsal pathway. The ventral pathway takes major part of the signals from the P cells in the retina, via the P channel in LGN, and goes through the cortical regions V1, V2, V4, PIT, CIT, AIT. The dorsal pathway takes major part of the signals from the M cells in the retina, via the M channel in LGN and goes through the cortical regions V1, V2, MT, LIP, MST, VIP, 2a and further on. It was suggested (e.g., Kandel et al. 1994 [1]) that the ventral pathway is mainly responsible for the form and colomanipulater of objects while the dorsal pathway mainly processes information about motion and location. In other words, the ventral pathway is a "what" pathway and the dorsal pathway a "where" pathway. Fig. 1 shows the nature of the processing along the "where" and "what" pathways, both of which are shaped by not only sensory inputs but also the motor outputs.

* Corresponding author.
E-mail addresses: jizhengp@gmail.com (Z. Ji), weng@cse.msu.edu (J. Weng).

http://dx.doi.org/10.1016/j.robot.2015.03.004
0921-8890/

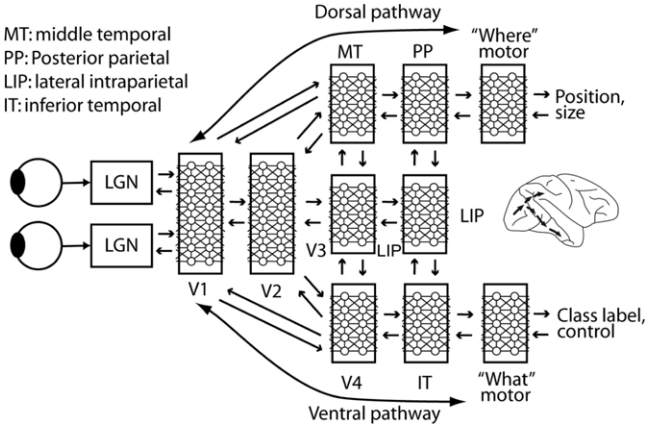

Fig. 1. The neurological nature of the processing in the "where" and "what" pathways.

The "where" and "what" pathways are also shown interconnected, revealing that the functions of attention and recognition perform interactively, known as a chicken-and-egg problem. Without attention, recognition cannot do well: recognition requires attended areas for the further processing. Without recognition, attention is limited: attention does not only need bottom-up saliency-based cues, but also top-down object-dependent and location-dependent signals. The successful modeling of the "where" and "what" pathways thus involves the integration of bottom-up and top-down cues to provide coherent control signals for the interplay between attentional tuning and object recognition.

1.1. Attention control

Three types of attention controls are well accepted and widely discussed: (a) the *bottom-up attention*, varying with an external world change, (b) the *location-based top-down attention*, derived from imposed signals representing locations and (c) the *object-based top-down attention*, derived from imposed signals representing object identities.

Bottom-up attention has been modeled by a variety of hand-designed features from visual inputs, especially in the aspect of saliency properties. One of the earliest theoretical studies to address this problem is a psychophysical literature. Treisman et al. [2] proposed that the brain develops feature maps in different cortical regions and combine them together via the processing of a master mapping. A more explicit computational model of bottom-up attention was introduced by Koch and Ullman in 1985 [3], in which a "saliency map" was used to encode stimuli saliency at every location in the visual scene. Elementary features, such as color, orientation, direction of movement and disparity are represented parallelly in different topographical maps, corresponding to the feature maps of Treisman's model. A selective mapping transforms these representations into a central representation, where inhibition of returns suppressed the current attended location and enabled the shifting to the next salient location. More recently, Itti and Koch et al. 1998 [4] constructed Gaussian pyramids to extract basic intensity and color features from red–green opponency channels and blue–yellow opponency channels. Four orientation-selective pyramids were further generated using Gabor filtering at 0, 45, 90, and 135 degrees. In total, 42 feature maps (6 for intensity, 12 for color, and 24 for orientation) were created and combined in a normalized scale for a saliency map. Backer et al. [5] applied the similar strategy above to an active vision system, called NAVIS (Neural Active VISion), emphasizing the visual attention selection in a dynamic visual scene. Instead of directly using some low level features like orientation and intensity, they accommodated additional processing to find mid-level features, e.g., symmetry and eccentricity, to build the feature map. Fusion of conspicuity maps was conduced using what is called Dynamic Neural Fields [6].

In contrast to aforementioned bottom-up attention models purely driven by visual inputs, volitional shifts of attention are also thought to be performed top-down, through control of high-level concepts (i.e., spatio-defined and object-dependent). Early top-down attention models selected the conspicuous locations regardless of being occupied by objects or not, as is called location-based top-down control. Olshausen et al. 1993 [7] proposed a model of how visual attention could solve the object recognition problem of location and scale invariance. A representative top-down attention model was discussed later by Tsotsos et al. 1995 [8], who implemented attention selection using a combination of a bottom-up feature extraction scheme and a top-down selective tuning scheme. A more extreme view was expressed by the "scanpath theory" of Stark and Choi 1996 [9], in which the control of eye movements was almost exclusively under top-down control. Mozer et al. 1996 proposed a model called MORSEL [10], to combine the object recognition and attention, in which attention was shown to help recognition. A top-down, knowledge-based recognition component, presented by a hierarchical knowledge tree, was introduced by Schill et al. 2001 [11], where object classes were defined by several critical points and the corresponding eye movement commands that maximized the information gain. Rao et al. 2004 [12] described an approach allowing a pair of cooperating neural networks to estimate object identity and object transformations, respectively.

1.2. Integration of attention and recognition

Aforementioned work provided computational models of attention (both bottom-up and top-down) and its link to object recognition capabilities. However, limited work has addressed an issue on their overall interactions and integrations. Specifically, what is the computational causality that can account for the concurrent development of the "where" and "what" pathways by both bottom-up and top-down controls? Deco and Rolls 2004 [13] presented a model to integrate both invariant object recognition and top-down attentional mechanisms on a hierarchically organized set of visual cortical areas. The model displayed location-based and object-based covert visual search by using attentional top-down feedback. Unfortunately the proposed architecture was not demonstrated in a scalable network for an engineering performance of attention and recognition. Moreover, top-down connections were used to propagate top-down signals only, without any internal development through neural computations. Our recent Where–What Network 1 (WWN-1) (Ji et al. [14]) is a developmental architecture for an interactive integration of top-down attention (both location-based and object-based) and recognition. Rather than the simulations of fMRI data, the engineering performance of recognition rate and attended spatial locations are presented in the experiment. However, the bottom-up feature-based attention is missing in the network, and limited complexity of "where" and "what" outputs (5 objects and 5 locations) looses its generality and scalability.

To solve the limitations of existing mechanisms in modeling interaction of attention and recognition, along with a demonstration of engineering-grade performance, we propose a new developmental network, called Where–What Network 2 (WWN-2). The proposed work is the neuromorphic developmental model for a *sensorimotor system*, interactively incorporating bottom-up attention, top-down attention and object recognition as *emergent internal actions*, and solving the where–what problems for *naturally unsegmented inputs* with *low computational complexity*. The highlighted 4 components are considered as critical requirements for