ARTICLE IN PRESS

Biologically Inspired Cognitive Architectures xxx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Biologically Inspired Cognitive Architectures

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

Research article

Towards a model of visual recognition based on neurosciences

Adrían González-Casillas*, Luis Parra*, Luis Martin*, Cynthia Avila-Contreras,
Raymundo Ramirez-Pedraza, Natividad Vargas, Juan Luis del Valle-Padilla, Félix Ramos*

Department of Computer Science, Center for Research an Advanced Studies of the National Polytechnic Institute (CINVESTAV IPN) Unidad Guadalajara, Guadalajara, Jalisco, Mexico

ARTICLE INFO

Keywords:
Brain model
Visual features
Perception
Cognitive architecture
Bioinspired model
Visual memory

ABSTRACT

Cognitive sciences and computer vision have proposed diverse models to acquire, transform and interpret visual information, mainly aimed to achieve realistic, yet efficient approaches to those capacities. One of the key aspects of visual processing is the identification of objects in the scene, that entails the perceptual association of visual features with semantic information extracted from memory. In this study, we present a model for visual recognition that resembles the way the human's brain interacts to achieve this process. The model describes the processes in V1 and V2 to extract features of lines, angles, and contours; as well as a template matching process in ITC, that uses early low spatial frequency visual information to bias the available comparisons. Operations of prefrontal areas DLPFC and VLPFC to maintain the representation and OFC to give a response are also described. Our proposal is intended to be the basis to treat visual information in a broader cognitive architecture. We find that matching of ITC templates provide a general and biologically inspired representation for objects. We also show how the use of low spatial frequency visual information can lead to a faster identification process when previous data exists. This is achieved by selecting a small number of ITC templates to handle the incoming bottom-up input.

Introduction

In optimal conditions, vision is the main source of information from the environment, therefore, it is the most studied sensory system and crucial to understanding human perception.

Visual processing involves mechanisms to generate internal abstract representations, by applying multiple transformations to the light of environmental objects that reaches photoreceptors in the eye. Recognition refers to giving a meaning to such representations (Albright, 2015, chap. 28), regardless of simplicity, and it is shaped by the current sensory activations, past sensory experiences and associations between these experiences.

Effective and efficient visual recognition is critical in various scenarios, like detecting dangerous predators hidden in the woods or interpreting a red traffic light while driving. Visual recognition plays an important role in setting basic information required to generate plans to interact with the environment, and then be able to make decisions over possible actions to satisfy goals.

Russell and Norvig (2009) state some commonly required properties that a general artificial intelligence should include, such as being capable of sensing, perceiving, learning, representing knowledge, and making decisions. The issue is often how all these capabilities coexist in the same schema. *Cognitive Architectures (CA)* are useful approaches to construct this type of systems, because they aim to describe the structure and interactions of the human mind's functions, and how to integrate them.

The main motivation of this work is to build a model of visual processing for virtual entities that resemble the way humans do and contribute to a better comprehension of the mechanisms and functions involved in the process of visual object recognition tasks. The emphasis is on bottom-up and top-down, as well as the process importance when encompassed in a larger a cognitive system, such as a cognitive architecture.

In this paper, we present a cognitive model for visual processing and object recognition that can be integrated with a broader cognitive architecture, by setting the basis of the different processes and brain areas involved. This model has modules associated with brain areas that perform operations of one or various *Cognitive Functions (CF)*. The CF provide specific human-like capabilities to the overall CA in which these are integrated.

* Corresponding authors.
E-mail addresses: augonzalez@gdl.cinvestav.mx (A. González-Casillas), laparra@gdl.cinvestav.mx (L. Parra), ldmartin@gdl.cinvestav.mx (L. Martin), framos@gdl.cinvestav.mx (F. Ramos).

https://doi.org/10.1016/j.bica.2018.07.018
Received 9 May 2018; Received in revised form 13 July 2018; Accepted 14 July 2018

We distinguish between two main cognitive functions that compound visual processing: sensory system and perception. The first, sensory system, is linked to the pure sensory aspects of the stimulus, that encompasses neural activations of visual features elicited by stimuli. Those characteristics could be as simple as the orientation of bars, or more complex as contour integration (Gilbert, 2015, chap. 25). We propose operations to extract those features in a similar way to the visual cortex. On the other hand, perception is related to the integration of features into a particular entity, it concerns the representation and discrimination of visual objects which leads to visual recognition (Miyashita, 1993). In that aspect, the model introduces a function to associate visual representations to previously presented visual objects.

Memory participates in the retrieval and maintenance of visual representations generated by perception. We divide memory into two functions: working memory, in the model it keeps the task information and visual representation; and semantic memory, that has information about known objects, but it is beyond the scope of the present study.

This paper is organized as follows: In Section 'Related models of visual processing with recognition' we describe some other bioinspired models for visual recognition. Section 'Neuroscientific evidence' explains the neuroscientific basis of the proposal. Section 'Biomodel of visual recognition' presents the proposed model and describes in detail the different processes considered. The study case used to evaluate the model and the obtained results are described in Section 'Experimentation and results'. Section 'Discussion and conclusions' presents a discussion about the limitations of the model and future work.

Related models of visual processing with recognition

The integration of biological research fields, as neurosciences, in hand of engineering and computer sciences has provided a vast amount of evidence on psychophysics and neural aspects of the visual system (Cox & Dean, 2014), which has allowed the development of diverse models of artificial vision.

Neural Networks (NN) are a common approach of many biologically inspired computational models for visual processing and recognition.

VisNet (Rolls, 2012) is a significant model for view-invariant object recognition that is biologically plausible and uses NN. The model makes a correspondence between layers of the network and some areas in visual cortex (V2, V4) and temporal visual cortex (inferior and anterior), thus building a hierarchy of the visual ventral pathway. It implements competitive learning to develop conjunctions of features, complemented by a temporal trace, by spatial continuity or both. Due to simplification, it does not include early-level visual processing areas, like the retina, thalamus or V1. VisNet is comparable to the HMAX model (Riesenhuber & Poggio, 1999a), that has a hierarchical structure to process visual features (orientation) as well but uses a multiscale pyramid of the image to extract them on each position at different levels, and combines them to form prototypes that will be integrated into an image dictionary. According to Born, Galeazzi, and Stringer (2017), the difference between these two systems is that HMAX needs more computational units than VisNet and uses an external non-biologically inspired classifier at its output layer, while the other encodes the classification itself, which is more likely to occur in the brain in a similar way. We find adequate the general hierarchical processes of these two systems, but we think low-level processes should be considered as well.

A widely used approach to object recognition are Convolutional Neural Networks (CNN), in which convolution masks resemble the receptive fields of the human visual cortex (Liu, Fang, Zhao, Wang, & Zhang, 2015). A traditional CNN is divided into 2 parts: feature extraction and classification. In feature extraction, several stages of filtering and pooling are performed, as the information advances in the layers, it becomes invariant to the position and scale (Nielsen, 2018). The filters used for feature extraction are learned from the images used in the training of the network, and the filters at higher levels are formed by linear combinations from the filters at lower levels, like how the human visual system reacts to more complex characteristics as the information flows from the bottom up. When sufficient features are extracted, the classification process is performed by a fully connected NN, resulting in the activation of certain classes in the output layer corresponding to the type of object detected (Nielsen, 2018).

Another type of biologically inspired approaches are the ones that not necessarily imitate the behavior at a neural level but take the functions suggested by neurosciences in distinct parts of the brain. This helps to simplify the model, not worrying about the neural interactions, but the overall actions instead. An example of this type of systems is NVRS (Khosla, Huber, & Kanan, 2014), which integrates two modules (attention and object recognition) to detect and localize objects and then classify them into one of several pre-defined object classes. These modules perform algorithms inspired in visual processing descriptions presented in cognitive sciences. Despite its strong biological background, it assumes that the processes of one module are performed before the ones of the second module, but some of them are in fact executed simultaneously in the brain, for example, feature saliency and feature extraction.

The CA aim to explain a broader theory of how the mind works, so they tend to go further than just describing the acquisition of sensory information. ACT-R and SOAR are both widely known architectures that work with production rules. These rules are executed during a cognitive cycle directed by procedural memory. The procedural memory evaluates information stored in working memory to execute an action, also, the working memory content can be updated and initiate responses or more cycles (Laird, Lebiere, & Rosenbloom, 2017). Specifically, ACT-R has a vision module divided into three buffers: a visual buffer that maintains the representation of an object; a visual-location buffer that preserves the location of an object; and a visual-state buffer that holds the internal state of the vision module (Anderson et al., 2004). Some ACT-R implementations integrate an attention module to pre-select information (Nyamsuren & Taatgen, 2013). On the other hand, SOAR uses graphs to make representations of scenes, objects, and their properties. However, none of the two architectures specifies the way in which they acquire and construct the object as theoretic evidence in neurosciences suggests (Gilbert, 2015, chap. 25), because they were not initially developed as biologically inspired.

At the end of this review, we could say that despite the advantages observed in the different models of visual recognition, the problem is that they are working at different levels, still not compatible: while those based on NN are useful due to their computational performance and biological acceptance, it is difficult to model a broad cognitive system (with diverse functions) that relies only on such an approach; while those based on cognitive architectures do not specify operations at the sensory level and they lack important details about the transformation of information. Based on that, we are trying to avoid the use of large networks that tend to be complex to build and arduous to maintain, so we propose a modular system in which each component computes specific operations based on evidence on the visual hierarchy, rescinding those of convolutional networks. In this way, we propose the convenience of feasible computational operations in independent components, and the description of a general cognitive system (in this case of visual recognition), which can be refined with the addition of more components and operations (of the same or another function).

Neuroscientific evidence

As many other tasks, visual processing requires the interaction of several brain structures. It is widely known that cortical visual system has two distinct pathways: identification and recognition of visual information are performed by activations of neurons along the ventral