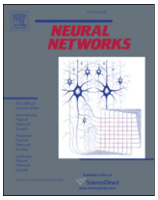
Neural Networks 21 (2008) 562–576

Contents lists available at ScienceDirect

Neural Networks

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



2008 Special Issue

Learning to recognize objects on the fly: A neurally based dynamic field approach

Christian Faubel*, Gregor Schöner

Institut für Neuroinformatik, Ruhr-Universität Bochum, Germany

ARTICLE INFO

Article history:
Received 6 December 2006
Received in revised form
7 March 2008
Accepted 7 March 2008

Keywords:
Object recognition
Computer vision
Robotics
Online learning
Human–machine interaction
Scene representation
Dynamic field theory
Binding

ABSTRACT

Autonomous robots interacting with human users need to build and continuously update scene representations. This entails the problem of rapidly learning to recognize new objects under user guidance. Based on analogies with human visual working memory, we propose a dynamical field architecture, in which localized peaks of activation represent objects over a small number of simple feature dimensions. Learning consists of laying down memory traces of such peaks. We implement the dynamical field model on a service robot and demonstrate how it learns 30 objects from a very small number of views (about 5 per object are sufficient). We also illustrate how properties of feature binding emerge from this framework.

1. Introduction

Autonomous robots generate flexible behavior based on their own sensory information. One defining feature of autonomous robots is that they operate in "natural environments", that is, environments that are not specifically designed for the robots' operation, are not metrically calibrated, and may change over time. Natural environments are difficult to model, requiring non-conventional engineering approaches such as learning from experience. One reason why autonomous robots need to be able to deal with such environments is that they are often expected to share environments with human users. Typical scenarios involving robot–human interaction are service robots, production assistants, support robotics in care or clinical settings, as well as entertainment robotics and robotic interfaces to information services. In many of these cases, the human user will be directing the attention and action of the robot onto objects in the shared environment. This requires the robot to perceive these objects and to be able to understand commands referring to the objects as well as to communicate about the objects.

Building scene representations is thus a central task that must be solved in most autonomous robotics scenarios, and certainly in all robotic scenarios involving goal-directed interaction with human users. To build scene representations that support interaction with human users, relevant objects must be segmented and recognized, and pose parameters must be estimated. Recognition involves associating objects with labels that can be used in communication with the human user.

Recognition in the context of scene representation differs from the general object recognition problem of computer vision. That latter problem is still largely unsolved, in particular, when objects are embedded in natural environments. Recognizing objects within robotic scenes is both a more difficult and a much simpler problem than general object recognition in computer vision. The added difficulty comes from the requirement that new objects are learned from a very small number of views, ideally a single view, typically a handful of views. The human user will teach new objects to the robotic system by directing attention to a part of the scene (e.g., through pointing) and labeling the object. The user may also give corrective input when the robot system tries to recognize the object in new poses. The number of times a human user is willing to provide such teaching signals is very limited, however. Another aspect of this difficulty is the required flexibility in which a single label may refer to different objects at different moments in time, such as when a new exemplar of a category of objects is being referred to. The robotic recognition system must be able to update object representations, replacing no longer relevant information with new input on the same fast timescale that human users find tolerable.

These difficulties are off-set by the fact that in typical robotic scenarios only a quite limited number of objects is relevant for meaningful interaction and object-directed action within a scene. All action within the scene can be used to update

* Corresponding author.
E-mail addresses: Christian.Faubel@neuroinformatik.rub.de (C. Faubel), Gregor.Schoener@neuroinformatik.rub.de (G. Schöner).

0893-6080/$ – see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.neunet.2008.03.007

object representations. User feedback may be available throughout operation, so that a small error rate is tolerable. Moreover, many robotic scenarios provide relatively simple viewing conditions. Objects to which a robot's attention or action is drawn may typically be placed so that the robot has them in view. Scenes will not be excessively cluttered with objects, occlusions will be limited in extent, recognition does not necessarily have to be successful from only limited component views of the object. The robotic system may also be endowed with a priori knowledge about learned objects. For instance, the approximate visual size of objects may be inferred from the learning trial and the position of the segmented object in the visual scene.

Humans themselves are, of course, extremely efficient in this category of visual tasks. In fact, our visual perception of our surroundings is, in large part, a form of scene representation in which an inventory of action-relevant items or items which had been closely examined are represented in detail, while visual details of the larger environment are much less reliably retained. A dramatic demonstration of this cognitive nature of scene representation comes from failures to detect change when visual transients are masked and the task directs attention at non-changing parts of the visual array (O'Regan, Rensink, & Clark, 1999).

This human capacity may be one motivation to search for neuronally inspired solutions to the problem of building scene representations. A look at the psychophysics of scene perception is, in fact, helpful in more precisely defining the problem. There are two limit cases that touch upon the problem of building scene representations and have been extensively studied both experimentally and theoretically. The first limit case involves visual working memory as a basis for making judgments about visual scenes (Henderson & Hollingworth, 1999). Such judgments may involve the detection of a particular target object in visual search, the detection of change in discrimination paradigms, or the estimation of object features. Some of these operations involve visual working memory characterized by limited capacity and a temporal contiguity constraint, so that new objects interfere with objects previously held in working memory. Set effects, the influence of context and, relatedly, the role of reference frames point, however, to the longer-term factors (Baddeley, 1986; Fuster, 1995).

Theoretical accounts for this form of scene representation is based on the notion of feature dimensions, a small number of which is required to characterize each object (Treisman, 1998). Cortical feature maps form the neurophysiological backdrop for this conception. These maps are assumed to separately represent different feature dimensions such as orientation, spatial frequency, or color. When an object is segmented and brought into the foreground (in the language of this field, when attention is focused on the object), the feature values along the different feature dimensions are bound into an "object file". This idea accounts for how tasks involving conjunctions of different feature dimensions differ through "feature binding" from tasks that can be solved on the basis of any individual feature dimension. How such binding in an object file would occur in neural terms is not quite clear. An alternative account postulates that there are specific neuronal mechanisms for binding, involving, for instance, correlation between neural spike trains (Malsburg, 1981; Raffone & Wolters, 2001). The neurophysiological reality of such a binding mechanism as well as its functional effectiveness are debated (see, e.g., the special issue of Neuron in 1999 (Volume 24, pages 7–125)).

The second limit case is object recognition on the basis of object categories learned over much longer timescales. In this work, the fundamental tension is between two requirements (Riesenhuber & Poggio, 1999; Serre, Wolf, & Poggio, 2005). The first, selectivity, is the capacity to discriminate between objects whose images are highly correlated. The most dramatic example is probably human face recognition: the images formed by faces are very similar across different faces, especially if compared to other kinds of visual objects, but humans are particularly astute at discriminating different faces. The second requirement is invariance of recognition under a broad set of image transformations that enables humans to recognize objects from different viewing angles, under partial occlusion, and at variable visual distances. A neurophysiologically based approach to this problem (Riesenhuber & Poggio, 2000) treats this form of object recognition as a largely feedfoward computational problem, in which complex features are extracted by neurons with relatively low spatial resolution. Pooling and a hierarchical organization of such feature detectors lead to inputs from which the winner category can be determined. In this framework, the problem of binding different features belonging to the same object does not arise as a separate problem. Complex features, in a sense, already bind values along any elementary feature dimension (Riesenhuber & Poggio, 1999). Alternatively, neuronal interaction may contribute to binding within such a feedforward architecture (Wersing, Steil, & Ritter, 2001).

When humans perceive and operate in a scene, they perform both forms of object recognition. The particular exemplars of object categories are perceived and memorized in their particular rendering and pose (Henderson & Hollingworth, 1999). But humans also recognize object categories, and may use such categorical perception to structure discourse about the objects in a scene. In fact, their perceptual categories influence the representations on which such visual working memory and visual discrimination is based (Schyns, Goldstone, & Thibaut, 1998).

The problem an autonomous robot must solve when it interacts with a human user in a given environment is somewhere halfway between these two limit cases. Working memory is a first step toward building scene representations, but is too fragile and short term to achieve that building by itself. Some longer-term maintenance of acquired knowledge about objects is required. That may still fall short of the extensive learning involved in acquiring new object categories for invariant recognition.

We propose that Dynamical Field Theory is a framework, within which a neural approach to working memory may be extended to endow representations with the longer-term stability required for scene representations, while at the same time remaining close to ongoing sensory input and providing the flexibility and fast learning capabilities needed to maintain and update scene representations. Dynamic Field Theory is a neuronally based theoretical approach to understanding embodied cognition. Originally developed to understand how movements are prepared (Bastian, Schöner, & Riehle, 2003; Erlhagen & Schöner, 2002), the ideas have been applied to a wide range of behaviors ranging from perception (Giese, 1999; Hock, Schöner, & Giese, 2003) to spatial memory (Schutte, Spencer, & Schöner, 2003).

Dynamic fields are distributions of neuronal activation defined directly over relevant perceptual or motor parameters (e.g. feature dimensions or movement parameters) rather than over the cortical surface. Conceptually, they are neuronal networks, in which the discrete sampling by individual neurons is replaced by a continuous neural field that represents the metric structure of the represented dimensions. Localized peaks of activation are units of representation. When the activation level in the peaks exceed a threshold (conventionally chosen to be zero), such peaks represent perceptual or motor decisions, both in the sense of detection decisions and in the sense of selection among competing inputs. The location of such peaks along the feature or motor dimension represents the outcome of estimation processes and encodes metric information about stimuli or motor states. The neuronal dynamics of such activation fields is governed both by inputs and by neuronal interaction, which stabilizes localized