Neurocomputing 205 (2016) 382–392

Contents lists available at ScienceDirect

Neurocomputing

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

Bio-inspired unsupervised learning of visual features leads to robust invariant object recognition

Saeed Reza Kheradpisheh [a,e], Mohammad Ganjtabesh [a,*], Timothée Masquelier [b,c,d,e]

[a] Department of Computer Science, School of Mathematics, Statistics, and Computer Science, University of Tehran, Tehran, Iran
[b] INSERM, U968, Paris F-75012, France
[c] Sorbonne Universités, UPMC Univ Paris 06, UMR-S 968, Institut de la Vision, Paris F-75012, France
[d] CNRS, UMR-7210, Paris F-75012, France
[e] CERCO UMR 5549, CNRS – Université de Toulouse, F-31300, France

ARTICLE INFO

Article history:
Received 17 October 2015
Received in revised form
17 March 2016
Accepted 18 April 2016
Communicated by Mingli Song
Available online 10 May 2016

Keywords:
View-Invariant object recognition
Visual cortex
STDP
Spiking neurons
Temporal coding

ABSTRACT

Retinal image of surrounding objects varies tremendously due to the changes in position, size, pose, illumination condition, background context, occlusion, noise, and non-rigid deformations. But despite these huge variations, our visual system is able to invariantly recognize any object in just a fraction of a second. To date, various computational models have been proposed to mimic the hierarchical processing of the ventral visual pathway, with limited success. Here, we show that the association of both biologically inspired network architecture and learning rule significantly improves the models' performance when facing challenging invariant object recognition problems. Our model is an asynchronous feedforward spiking neural network. When the network is presented with natural images, the neurons in the entry layers detect edges, and the most activated ones fire first, while neurons in higher layers are equipped with spike timing-dependent plasticity. These neurons progressively become selective to intermediate complexity visual features appropriate for object categorization. The model is evaluated on 3D-Object and *ETH*-80 datasets which are two benchmarks for invariant object recognition, and is shown to outperform state-of-the-art models, including DeepConvNet and HMAX. This demonstrates its ability to accurately recognize different instances of multiple object classes even under various appearance conditions (different views, scales, tilts, and backgrounds). Several statistical analysis techniques are used to show that our model extracts class specific and highly informative features.

1. Introduction

Humans can effortlessly and rapidly recognize surrounding objects [1], despite the tremendous variations in the projection of each object on the retina [2] caused by various transformations such as changes in object position, size, pose, illumination condition and background context [3]. This invariant recognition is presumably handled through hierarchical processing in the so-called ventral pathway. Such hierarchical processing starts in V1 layers, which extract simple features such as bars and edges in different orientations [4], continues in intermediate layers such as V2 and V4, which are responsive to more complex features [5], and culminates in the inferior temporal cortex (IT), where the neurons are selective to object parts or whole objects [6]. By moving from the lower layers to the higher layers, the feature complexity, receptive field size and transformation invariance increase, in such a way that the IT neurons can invariantly represent the objects in a linearly separable manner [7,8].

Another amazing feature of the primates' visual system is its high processing speed. The first wave of image-driven neuronal responses in IT appears around 100 ms after the stimulus onset [1,3]. Recordings from monkey IT cortex have demonstrated that the first spikes (over a short time window of 12.5 ms), about 100 ms after the image presentation, carry accurate information about the nature of the visual stimulus [7]. Hence, ultra-rapid object recognition is presumably performed in a feedforward manner [3]. Moreover, although there exist various intra- and inter-area feedback connections in the visual cortex, some neurophysiological [9,10,3] and theoretical [11] studies have also suggested that the feedforward information is usually sufficient for invariant object categorization.

Appealed by the impressive speed and performance of the primates' visual system, computer vision scientists have long tried to "copy" it. So far, it is mostly the architecture of the visual system that has been mimicked. For instance, using hierarchical

* Corresponding author.
E-mail addresses: kheradpisheh@ut.ac.ir (S.R. Kheradpisheh), mgtabesh@ut.ac.ir (M. Ganjtabesh), timothee.masquelier@alum.mit.edu (T. Masquelier).

http://dx.doi.org/10.1016/j.neucom.2016.04.029
0925-2312/

feedforward networks with restricted receptive fields, like in the brain, has been proven useful [12–17]. In comparison, the way that biological visual systems learn the appropriate features has attracted much less attention. All the above-mentioned approaches somehow use non biologically plausible learning rules. Yet the ability of the visual cortex to wire itself, mostly in an unsupervised manner, is remarkable [18,19].

Here, we propose that adding bio-inspired learning to bio-inspired architectures could improve the models' behavior. To this end, we focused on a particular form of synaptic plasticity known as spike timing-dependent plasticity (STDP), which has been observed in the mammalian visual cortex [20,21]. Briefly, STDP reinforces the connections with afferents that significantly contributed to make a neuron fire, while it depresses the others [22]. A recent psychophysical study provided some indirect evidence for this form of plasticity in the human visual cortex [23].

In an earlier study [24], it is shown that a combination of a temporal coding scheme – where in the entry layer of a spiking neural network the most strongly activated neurons fire first – with STDP leads to a situation where neurons in higher visual areas will gradually become selective to complex visual features in an unsupervised manner. These features are both salient and consistently present in the inputs. Furthermore, as learning progresses, the neurons' responses rapidly accelerates. These responses can then be fed to a classifier to do a categorization task.

In this study, we show that such an approach strongly outperforms state-of-the-art computer vision algorithms on view-invariant object recognition benchmark tasks including 3D-Object [25,26] and ETH-80 [27] datasets. These datasets contain natural and unsegmented images, where objects have large variations in scale, viewpoint, and tilt, which makes their recognition hard [28], and probably out of reach for most of the other bio-inspired models [29,30]. Yet our algorithm generalizes surprisingly well, even when "simple classifiers" are used, because STDP naturally extracts features that are class specific. This point was further confirmed using mutual information [31] and representational dissimilarity matrix (RDM) [32]. Moreover, the distribution of objects in the obtained feature space was analyzed using hierarchical clustering [33], and objects of the same category tended to cluster together.

2. Materials and methods

The algorithm we used here is a scaled-up version of the one presented in [24]. Essentially, many more C2 features and iterations were used. Our code is available upon request. We used a five-layer hierarchical network $S_1 \rightarrow C_1 \rightarrow S_2 \rightarrow C_2 \rightarrow classifier$, largely inspired by the HMAX model [14] (see Fig. 1). Specifically, we alternated simple cells that gain selectivity through a sum operation, and complex cells that gain shift and scale invariance through a max operation. However, our network uses spiking neurons and operates in the temporal domain: when presented with an image, the first layer's S_1 cells, detect oriented edges and the more strongly a cell is stimulated the earlier it fires. These S_1 spikes are then propagated asynchronously through the

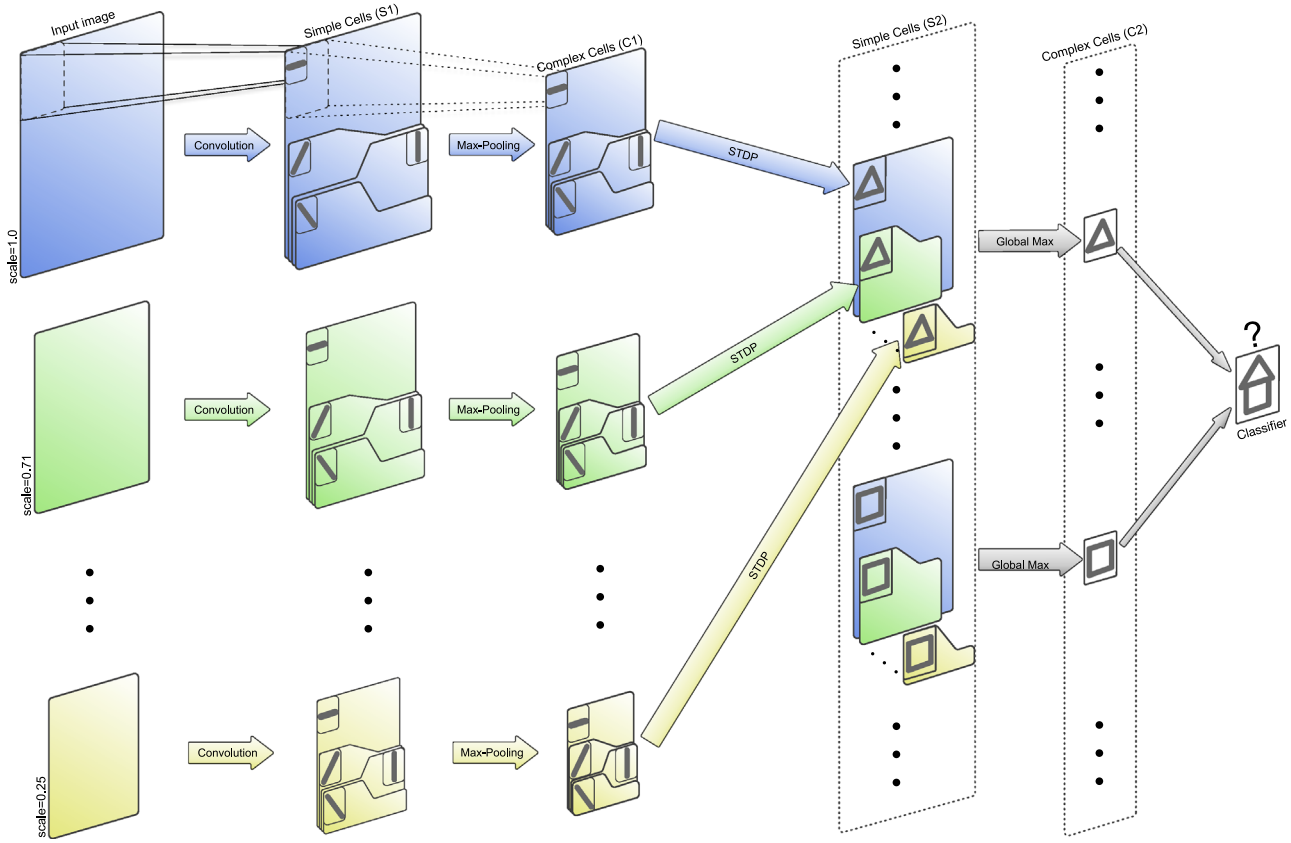

Fig. 1. Overview of our 5 layered feedforward spiking neural network. The network processes the input image in a multi-scale form, each processing scale is shown with a different color. Cells are organized in retinotopic maps until the S_2 layer (included). S_1 cells of each processing scale detect edges from the corresponding scaled image. C_1 maps sub-sample the corresponding S_1 maps by taking the maximum response over a square neighborhood. S_2 cells are selective to intermediate complexity visual features, defined as a combination of oriented edges of a same scale(here we symbolically represented a triangle detector and a square detector). There is one S_1–C_1–S_2 pathway for each processing scale. Then C_2 cells take the maximum response of S_2 cells over all positions and scales and are thus shift and scale invariant. Finally, a classification is done based on the C_2 cells' responses (here we symbolically represented a house/non-house classifier). C_1 to S_2 synaptic connections are learned with STDP, in an unsupervised manner. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)