Neurocomputing 000 (2018) 1–12

Contents lists available at ScienceDirect

Neurocomputing

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

Object recognition in clutter color images using Hierarchical Temporal Memory combined with salient-region detection

Radoslav Škoviera [a], Ivan Bajla [b,*], Júlia Škovierová [a]

[a] Czech Institute of Informatics, Robotics, and Cybernetics, Czech Technical University in Prague, Zikova 1903/4, Prague 6, 166 36, Dejvice, Czech Republic
[b] Institute of Measurement Science, Slovak Academy of Sciences, Dúbravská cesta 9, Bratislava 841 04, Slovak Republic

ARTICLE INFO

Article history:
Received 10 November 2017
Revised 12 February 2018
Accepted 9 April 2018
Available online xxx

Communicated by Yongdong Zhang

Keywords:
Artificial intelligence
Machine learning
Pattern recognition
Hierarchical Temporal Memory
Image saliency

ABSTRACT

The essential goal of this paper is to extend the functionality of the bio-inspired intelligent HTM (Hierarchical Temporal Memory) network towards two capabilities: (i) object recognition in color images, (ii) detection of multiple objects located in *clutter* color images. The former extension is based on the development of a novel scheme for the application of three parallel HTM networks that separately process color, texture, and shape information in color images. For the latter HTM extension, we propose a novel system in which HTM is combined with a modified model of computational visual attention. We adopt the results of Bi et al. (2010), Hu et al. (2005), and Kučerová (2011) and add new elements for the calculation of image saliency maps. The proposed algorithm enables one to automatically locate individual objects in clutter images. For computer experiments, a special image database is created to simulate ideal single object images and cluttered images with multiple objects on an inhomogeneous background. The recognition performance of HTM alone and in combination with the salient-region detection method is evaluated. We show that the attention subsystem is able to satisfactorily locate multiple objects in clutter color images with an inhomogeneous background. We also perform benchmark calculations for two selected computer vision methods used for object detection in color clutter images. Namely, the cascade detector and template matching methods are used. Our study confirms that the proposed attention system can improve the capabilities of HTM for object classification in cluttered images. The compound system of visual attention and HTM outperforms the compared methods in both criteria (recall and correct detection rate). However, as expected, the system cannot match the recognition accuracy achieved by HTM for single object images, and thus, further research is needed.

1. Introduction

The tremendous expansion of digital techniques for image acquisition and the ubiquity of web connections in recent decades have evoked the creation of an enormous number of image data bases that are accessible to web users. Searching/retrieving of images from such databases for various aims has become an everyday task in many application areas. Especially requested are the so-called content-based image retrieval (CBIR) techniques and systems [1–3]. In particular, for large image data sets, methods of mining semantic relations between unlabeled image data are needed. The authors of [4] proposed a novel topology preserving hashing approach that achieved better image search results than achieved in previous papers, e.g., [5–7]. Recently, Chenggang et al. [8] joined the hashing coding with a Deep Neural Network. Another area of complex event recognition methods, attracted the attention of researchers (e.g., [9]). These methods can serve to further extension of CBIR usage in web applications. Although a number of image-retrieval methods have been proposed and explored to date, no satisfactory general solutions are available. The core problems are as follows: (i) choice of suitable image features for image content representation, (ii) efficient image object detection/recognition in cluttered images that frequently occur in CBIR tasks. In previous studies [10,11] several aspects of solving these two problems by the application of a bio-inspired Hierarchical Temporal Memory (HTM) system have been addressed. This system can be considered as a type of intelligent network based on the deep learning principle [12]. Pal et al. [10] studied the use of a color histogram technique together with the HTM model for retrieving the final images post-classification of the query image. Škoviera and Bajla [11] focused on exploring the possibilities of the HTM network for application to CBIR when color features are used. Application of HTM to CBIR

* Corresponding author.
E-mail address: umerbajl@savba.sk (I. Bajla).

https://doi.org/10.1016/j.neucom.2018.04.030
0925-2312/

tasks has however revealed similar problems as outlined above, because the training images for HTM are characterized by single-centered objects that are usually placed to a homogeneous background.

It appears that to overcome the limitations of bio-inspired feed-forward models of the visual cortex, their application to multiple object recognition tasks requires the support of a computational model of visual attention. Several papers have been devoted to this research topic. A question was explored in [13] regarding to what extent saliency-based bottom-up attention can be used to extract useful information about the location, size, and shape of objects in cluttered scenes. The rigorous quantitative analysis of the authors showed the usefulness of the synergy between recognition and attention. Chikkerur et al. [14] proposed a two-stage approach to the recognition of objects in clutter. In the first stage, a Bayesian (feature-based) model of visual attention was used to isolate the target object while suppressing the surrounding clutter. The results were then fed into a hierarchical feed-forward model of object recognition in the ventral stream. The authors showed that attentive processing improves recognition in comparison to purely feed-forward processing. An improvement of feed-forward object recognition via a biologically plausible saliency mechanism was explored and demonstrated in the papers [15] and [16]. Recently two different models have been proposed in which saliency function is directly combined with a network for object class recognition. In [17] a salient hierarchical model for object recognition was proposed that is characterized by two contributing features: (1) a traditional saliency model is modified to achieve more robust saliency estimation, (2) this part is combined with the Hierarchical Maximization architecture (HMAX), the model of immediate object recognition in primate visual cortex.

To the best of our knowledge, the recent work of [18] represents the first attempt to combine the HTM network model with an image saliency approach. The authors proposed a supervised learning method for the recognition of objects in different orientations. Instead of conventional color image input vectors, the combined model exploits a preceding saliency detection step that isolates the region of interest, releasing the HTM learning procedure from redundant information. The proposed approach has several limitations that hinder from its application to colored clutter images containing many different objects located anywhere in the image. Therefore, the basic motivation of our research was to extend the possibilities of both parts of the combination of the HTM system and a visual attention system including saliency. We propose a novel version of such a combination with the following contributions: (1) the saliency of color images is calculated as a combination of a suppression map with contrast features; (2) a combined feature based on discriminative local regions is then calculated; (3) finally, instead of the standard HTM implementation with gray-level images as inputs, a system of three parallel HTM networks is proposed – each HTM network processes a separate image feature map, and a coupled *k*-NN classifier is proposed for weighted supervised classification of belief vectors inferred by individual HTM networks. Computer experiments incorporate the generation of multi-object images that simulate clutter images in a similar way as in [14] (placement of individual objects at random position and scale in a test multi-object image). In comparison to the above-mentioned combined models of feed-forward processing and image saliency maps [17,18], which are applied exclusively to images with one dominant object, we attempted to solve a more sophisticated multiple object recognition task in the clutter images (simulated models). The important goal of the paper was to validate the feasibility of one-level HTM network application to the task of object recognition in multi-object clutter images.

This paper is organized as follows. In Section 2, the basic description of a memory-prediction HTM network is provided. In addition, a novel system of parallel coupling of three HTM networks is proposed that processes three separate image features of edge, texture, and color. Section 3 is devoted to the description of the image data set construction needed for implementation of computer experiments with the proposed compound system. Two classes of images are used, namely, partially cluttered and fully cluttered images. In Section 4 the details of the proposal of a novel color image saliency model are explained. This is a key section in which the main contributions of our approach to the combination of the HTM network with the color image saliency mapping (ISM) model are presented. Section 5 is oriented towards the implementation of computer experiments. Two benchmark methods are selected and briefly described. The proposed compound system –(ISM + HTM) is compared to cascade detectors, and to the template matching approach. All three methods are applied to the identical testing images. In Section 6, the obtained results are presented and discussed. We used five different characteristics of object detection/recognition success: accuracy, overlap, recall, clutter, and hits. Finally, Section 7 presents the conclusions of our findings and discusses the contributions and limitations of the proposed approach presented herein. Unresolved issues in this domain necessitating further research are also outlined.

2. Hierarchical Temporal Memory

2.1. Basic description

The HTM is a memory-prediction network proposed in [19,20] and distributed initially by Numenta, Inc. as a free software package NuPIC [21,22]. Promising results in applications of HTM have been achieved especially in the field of visual object recognition/classification, e.g., [23–28]. It represents a hierarchical Bayesian network and can be assigned to the class of Deep Belief Networks in artificial intelligence [29,30], [12]. In more detail, it can be described as a hierarchy of several layers (levels) consisting of basic operational units called nodes.

The effective area from which a node receives its input is called the *field of view* or *receptive field* of the node. The individual levels are ordered in a hierarchical tree-like structure (see Fig. 1 as a prototypical HTM network). There is a zero sensory level of the HTM, which serves as an input to the first level of nodes. In our case, the zero level represents a visual field of image pixels or feature maps derived from it. At the top level, there is only one node that serves for classification. In this role, various classifiers can be used. Each HTM node works in two modes – learning and inference. In the learning mode the node performs two operations: spatial pooling and temporal pooling. Once these two steps are completed, the node is switched to inference mode. In practice, all nodes within one hierarchy level are considered to be equivalent. Since the use of smooth temporal dependencies of input spatial patterns is an essential characteristic of HTM, its learning process utilizes either native sequence of images (e.g., video captured by a camera) or (in case of static images) an artificially generated sequence of images using various exploration schemes.

In the first step of the learning process, the node memorizes the representative spatial patterns (coincidences) from its receptive field that results in creating a code-book of image patterns. After reaching the requested number of quantization centers, the memorization process is stopped. The ultimate goal of HTM learning is to detect correct invariant representations of the input world based on the temporal relationships contained in the learning sequence. To achieve this, one needs a frequency of transition events, i.e., co-occurrences of the memorized coincidences in adjacent time instances. A sequence of input patterns generates a sequence of n coincidences within the node. In HTM theory [20,21], the temporal relations are described in a form of the first-order Markov graph