Neurocomputing 133 (2014) 328–341

Contents lists available at ScienceDirect

Neurocomputing

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

Improving invariance in visual classification with biologically inspired mechanism

Tang Tang, Hong Qiao *

Chinese Academy of Sciences, Institute of Automation, State Key Laboratory of Management and Control for Complex Systems, Zhongguancun East Road 95#, Beijing, China

ARTICLE INFO

Article history:
Received 1 July 2013
Received in revised form
18 November 2013
Accepted 27 November 2013
Communicated by A. Belatreche
Available online 4 January 2014

Keywords:
Biologically inspired
Visual classification
Max-pooling
Template matching

ABSTRACT

A computational model of visual cortex has raised great interest in developing algorithms mimicking human visual systems. The max-operation is employed in the model to emulate the scale and position invariant responses of the visual cells. We further extend this idea to enhance the tolerance of visual classification against the general intra-class variability. A general architecture of the basic block constituting the model is first presented. The architecture adaptively chooses the best matching template from a set of competing templates to predict the label of the incoming sample. To optimize the non-convex and non-smooth objective function resulted, we develop an algorithm to train each template alternately. Experiments show that the proposed method significantly outperforms linear classifiers as a template matching method in several image classification tasks, and is much more computationally efficient than other commonly used non-linear classifiers. In the image classification task on the Caltech 101 database, the performance of the biologically inspired model is obviously boosted by incorporating the proposed method.

1. Introduction

Visual object classification has been a long-term interest in computer vision for its important role in a variety of applications, such as content based image retrieval, intelligent transportation systems, automatic video surveillance and human–computer interface. As a lot of works have been done to improve the related techniques such as local image descriptors [1], visual similarity metrics [3] and saliency detection [2], this field has made steady progress and adoption in some practical systems in the past few years. However, when challenged by the complex variation of visual objects in the real world, the performance of the artificial visual systems is far behind their biological counterparts.

The new findings in biology prompt investigators to take inspiration from biological visual systems. A model of the feed-forward pathway in visual cortex has been proposed to account for recognition ability of human [4,5]. In this model, template matching and max-pooling layers are alternately cascaded to construct an object representation hierarchy with increasing invariance against scale and translation transforms. This is actually an analog to ventral stream found in human visual cortex. While the redundant information from raw data is merged gradually along the hierarchy, the selective information is largely remained and enhanced by the repetitive matching process. This model provides a new insight into how the creatures cope with the complex visual world, and a series of experiments have been conducted to show the high coincidence of the model response with physiological data [5]. In a further evaluation of the model's utility, the visual recognition system combining the model with the commonly used classifiers exhibits state-of-the-art performance when compared with other functionally designed computer vision algorithms [6].

In light of the new path approaching human level visual systems revealed by the model, a great deal of works have been done to improve its performance and widen the applications. Mutch and Lowe [7] refined the model with several biologically plausible operations, which include sparse input organization, response inhibition, retaining some position information on the top level, and feature selection over final output. Meyers and Wolf proposed a new set of features for face processing derived from the model [8]. The first two stages of the extraction procedure are similar to the model, except for the center-surround processing added to the first matching layer. Instead of another two matching and pooling layers, the subsequent processing is replaced by a kernelized and regularized relevant component analysis and normalization to generate discriminative and localized features which are more suitable for face processing. Huang et al. proposed an enhanced model modified in two aspects including eliminating uninformative inputs under sparsity constraints and applying a boosting-based feature selection as a feedback loop [9]. Other variations of the model are also applied in gait recognition, digit recognition, scene classification, etc. [10–12].

While most of these researches seek to enhance the model by introducing more biologically plausible elements or task-related processing, the employment of the max-pooling mechanism is still

* Corresponding author.
E-mail address: hong.qiao@ia.ac.cn (H. Qiao).

0925-2312/$ - see front matter © 2014 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.neucom.2013.11.003

limited to provide the scale and translation tolerance in the low-level processing. However, research suggests that the max-pooling exists widely in biological visual systems across multi-levels in the visual recognition process, and plays a critical role in invariant response to mutable appearance of complex stimuli [13–15]. To extend the application of the max-pooling mechanism, in this paper we propose a novel architecture combining several templates based on the max-pooling mechanism and explore how to use it to cope with the general intra-class variability. The classification result is the summation of outputs from several template groups, in which competing templates responding optimally to different sub-classes are pooled over by max-operation.

Obviously the key to the effectiveness of the architecture is to find a proper set of templates, which can also be regarded as sub-classifiers. The linear SVM classifier is popular in many typical applications of computer vision, such as detection and recognition, for its high discrimination and low computational cost [16–18]. However, hardly any visual category can be completely distinguished from others by matching with just one template represented by a linear classifier. Fortunately, the proposed architecture can help to overcome this shortage while keeping its computational efficiency. In this paper, we mainly aim to design a training algorithm for the template ensemble with the proposed architecture. The training algorithm is based on minimizing the hinge loss function, which has been proved desirable for classification as used in SVM.

For the proposed architecture, it is easy to see that the classification of one sample only relies on those templates winning the competition within the groups. Conversely, one template only affects a part of the incoming samples. These two features unify the template training and data clustering into one process naturally, which allows us to train each template specifically for different sub-modes of data distributions. They also keep the specialization of templates in classification consistent with that in training. On the other hand, the use of hinge loss and max-pooling also brings the training process into trouble with the non-smooth and non-convex objective function. To address this issue, we first present how to train one template while keeping others fixed by searching descent direction on decomposed subspaces, and then extend it to train all the templates by optimizing each one alternately until no more decline in loss can be achieved.

In the experiment section, we show that a set of differentiated templates can be efficiently constructed by the proposed algorithm, and the accuracy of the resulting classifier is obviously higher than several classifiers commonly used for visual classification. On several face pattern and hand-written digit databases, the new architecture with a very few templates achieves the performance at the same level of the kernel-based SVM classifier with hundreds of support vectors, which means that the computational cost can be drastically reduced. In image classification task on the Caltech-101 database, the performance of biologically inspired model has been obviously improved by incorporating the proposed architecture.

The remainder of this paper is organized as follows. First we introduce the original standard model of the feed-forward pathway in visual cortex in Section 2, and then we explain the motivation of our method and give an overview of the method in Section 3. In Section 4, the training algorithm is described and discussed. In Section 5, we report results of the experiments on several visual image datasets. Finally we conclude this paper in Section 6.

2. The standard model of visual cortex and its biological inspiration

The standard model of visual cortex proposed in [4] is generally a feed-forward hierarchical model accounting for initial 100–200 ms in the visual processing of primate. The basic form of the model is comprised of four layers which are S1, C1, S2 and C2. The structure is illustrated in Fig. 1.

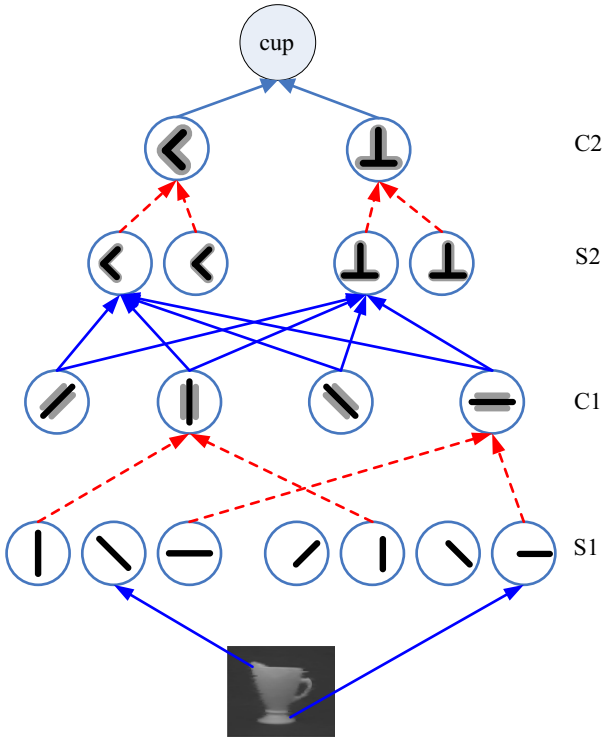

Fig. 1. The basic structure of visual cortex model in [4]. Red dashed lines indicate maximum operations and blue solid lines indicate matching operations. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this paper.)

In the S1 layer, a bank of Gabor filters in different scales, positions and orientations is applied to convolve with raw input image. The S1 units correspond to simple cells found in the primary visual cortex which are sensitive to bar shaped stimuli of specific orientations and scales.

Then in the C1 layer, outputs from the S1 units within a spatial neighborhood and across adjacent scales are merged by the maximum operation to produce responses with some tolerance against positions and scale variances. This layer is expected to imitate some complex cells in the primary visual cortex, which are widely accepted as bar detectors with larger receptive fields and less sensitive to shift and size than simple cells.

In the next S2 layer, matching is performed with more complex templates represented by combination of edges and bars in different orientations. The C1 units are expected to behave like view tuned neurons in the inferotemporal cortex. Biological experiments show that these neurons are highly selective to complex stimuli with a bell-shaped tuning curve over view variation. This property can be well modeled by the Gaussian kernel function measuring similarity between the C1 responses and the templates across all involved orientations. In [6], these templates are randomly extracted from class-specific or non-specific images.

In the C2 layer, the responses from the last layer are pooled with the maximum operation once again to produce the final responses. For each template stored in S2, only the best matching result within the whole image and over all scales is output. From the biological perspective, the C2 units are models of V4 neurons with large receptive fields, high invariance and selectiveness.

To classify images, the feature vector is then fed to simple linear classifiers, usually SVM or boosting, to obtain classification results. These classifiers are powerful enough to facilitate the features generated by the model. Additionally, recent work by