Neurocomputing 193 (2016) 155–166

Contents lists available at ScienceDirect

Neurocomputing

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

Dominant orientation patch matching for HMAX

Yan-Feng Lu [a], Hua-Zhen Zhang [b], Tae-Koo Kang [c,*], Myo-Taeg Lim [b,**]

[a] Institute of Automation, Chinese Academy of Sciences, Beijing, China
[b] School of Electrical Engineering, Korea University, Seoul, Republic of Korea
[c] Dept. of Information and Telecommunication Engineering, Sangmyung University, Cheonan, Republic of Korea

ARTICLE INFO

Article history:
Received 7 July 2015
Received in revised form
4 December 2015
Accepted 27 January 2016
Available online 23 February 2016

Keywords:
Object recognition
Classification
HMAX
Dominant orientation
Patch
Matching

ABSTRACT

The biologically inspired model for object recognition, Hierarchical Model and X (HMAX), has attracted considerable attention in recent years. HMAX is robust (i.e., shift- and scale-invariant), but it is sensitive to rotational deformation, which greatly limits its performance in object recognition. The main reason for this is that HMAX lacks an appropriate directional module against rotational deformation, thereby often leading to mismatch. To address this issue, we propose a novel patch-matching method for HMAX called Dominant Orientation Patch Matching (DOPM), which calculates the dominant orientation of the selected patches and implements patch-to-patch matching. In contrast to patch matching with the whole target image (second layer C1) in the conventional HMAX model, which involves huge amounts of redundant information in the feature representation, the DOPM-based HMAX model (D-HMAX) quantizes the C1 layer to patch sets with better distinctiveness, then realizes patch-to-patch matching based on the dominant orientation. To show the effectiveness of D-HMAX, we apply it to object categorization and conduct experiments on the CalTech101, CalTech05, GRAZ01, and GRAZ02 databases. Our experimental results demonstrate that D-HMAX outperforms conventional HMAX and is comparable to existing architectures that have a similar framework.

1. Introduction

Object recognition has been widely used in pedestrian detection, visual navigation for robots, and video surveillance [1–3]. In practical applications, the difficulties that arise in object recognition involve two conflicting requirements. On one hand, the recognition system needs to be specific and selective for various objects, which may generally look quite similar and share certain characteristic features (e.g., telling apart twins). On the other hand, the recognition needs to be tolerant or invariant to different appearance-altering transformations of an object (e.g., recognizing a person under disguise). Object variability in terms of rotation, illumination, and scale changes, especially in the case of cluttered backgrounds, seriously disrupts recognition [4,5]. For example, various pedestrian postures (e.g., running, squatting, standing, or stooping) in a cluttered environment make accurate recognition a challenging task. Many approaches have recently been proposed to address this issue.

Traditional appearance- and contour-based methods mainly use low-level visual features, such as textures, edges, and colors [6,7]. While these methods generally take these features into account, they do not selectively address discriminative features. They are sensitive to scale, shift, and rotation deformations and variations in illumination. Local feature-based methods combine local descriptors and keypoint detectors with spatial information. Representative local feature methods have been proposed, such as scale-invariant feature transform (SIFT) [5], speeded-up robust features (SURF) [8], gradient location and orientation histogram (GLOH) [9], and histogram of gradients (HOG) [10]. These methods are effective in terms of describing locally discriminative features, but they lack oriented local information. Bag-of-features [11] is effective for resolving this issue; however, the amount of structural information still falls short.

In recent years, significant advances have been made in the understanding of brain cognition in the biological vision field. The ventral pathway in the primate visual cortex, from the primary visual area V1 to the inferior temporal cortex IT, is thought to mediate object recognition, which is commonly called the “what pathway” [12]. Within the early visual areas along the pathway, such as V1, neurons tend to respond well to oriented bars or edges [13]. Neurons in the intermediate visual areas, such as V2 and V4, are no longer tuned to oriented bars only, but to other forms and shapes of intermediate complexity [14–17]. Finally, in high-visual

* Corresponding author.
** Corresponding author.
E-mail addresses: yanfeng.lv@ia.ac.cn (Y.-F. Lu), zhanghz@korea.ac.kr (H.-Z. Zhang), tkkang@smu.ac.kr (T.-K. Kang), mlim@korea.ac.kr (M.-T. Lim).

http://dx.doi.org/10.1016/j.neucom.2016.01.069
0925-2312/

areas, such as the inferior temporal cortex (IT), neurons are responsive to complex shapes, such as the image of a face or a hand [18–21]. These discoveries have provided biological support for early-stage psychophysical theories. Riesenhuber and Poggio developed these theories and presented an initial computational model of object recognition, called Hierarchical Model and X (HMAX), which attempts to model the recognition mechanism of the cortex [22]. Serre et al. improved the original HMAX significantly on a variety of large-scale, real-world object recognition databases. The extension of the original model is named the standard HMAX model [23], which demonstrates that the visual cognitive model efficiently enhances the performance of object recognition.

HMAX is a biologically inspired feature descriptor that focuses on feature invariance and selectivity. Although HMAX is robust to scale and position changes, the poor invariance to rotation has not been significantly improved [23]. The improvement of the rotation invariance of local features is a challenging issue, and some valid approaches have been proposed recently [24–26]. To improve robustness to rotation, HMAX simply introduces rotated versions of training images, which alone is very inadequate. The feature representation in the S2 layer of HMAX directly influences the overall performance of HMAX. In the conventional HMAX, the S2 layer measures the matching between a stored patch and the previous C1 layer at every position and scale. However, the matching does not take the directional information into account; in addition, the matching between the patch and the whole C1 layer is likely to bring in redundant information, which is prone to yielding poor feature discrimination. Due to these drawbacks, the overall performance of HMAX is limited. To solve this issue, we propose a new patch-matching method called Dominant Orientation Patch Matching (DOPM) and employ the DOPM in the HMAX structure for DOPM-based HMAX (D-HMAX). In contrast to the conventional HMAX, which directly measures the matching between a stored patch and the whole C1 layer, D-HMAX achieves a patch-to-patch matching based on the dominant orientations. We test the object categorization of D-HMAX and evaluate its accuracy in the CalTech101, CalTech5, GRAZ01, and GRAZ02 datasets.

The rest of this paper is organized as follows. In Section 2, we briefly review the standard HMAX model. In Section 3, we elaborate on the DOPM and DHMAX model. In Section 4, we present experimental results based on four public datasets. Finally, in Section 5, we give our conclusions.

2. HMAX review

The standard HMAX [23] is a general framework for the recognition of complex visual scenes, which is motivated by biology: it is a hierarchical structure that closely follows the mechanism of visual cortex and builds an increasingly complex and invariant feature representation by alternating between a maximum pooling and a patch matching operation. The HMAX is with four layers: S1, C1, S2, and C2, as shown in Fig. 1. We briefly describe the operations of each layer of the HMAX model as follows.

S1 layers : The units in the S1 layers correspond to simple cells in V1. The S1 units take the form of Gabor functions, which have been described as a model of cortical simple-cell RFs [27]. Gabor functions are defined as

$$G(x,y)=\exp\left(-\frac{{x_o}^2+\gamma^2{y_o}^2}{2\sigma^2}\right)\times\cos\left(\frac{2\pi x_o}{\lambda}\right),$$

$$\text{s.t}\quad x_o=x\cos\theta+y\sin\theta\quad\text{and}\quad y_o=-x\sin\theta+y\cos\theta, \tag{1}$$

where θ represents the orientation, λ is the wavelength, σ is the standard deviation, and γ indicates the spatial aspect ratio.

Given an input image, the S1 layer with orientation θ and standard deviation σ is calculated by

$$S1_{\sigma,\theta}=|G_{\sigma,\theta}*I|, \tag{2}$$

where $*$ denotes the convolution operation, I is the input image, and $G_{\sigma,\theta}$ is the Gabor function with specific parameters.

C1 layers : The C1 layers describe the complex cells in V1. The layers are the dimension reduction of the S1 layers obtained by selecting the maximum over local spatial neighborhoods. The maximum pooling operation over local neighborhoods increases invariance (providing some robustness to shift and scale transformations).

S2 layers : The S2 layers describe the similarity between the C1 layers and stored patches in a Gaussian-like manner using Euclidean distance. The responses of the corresponding S2 layers are calculated by

$$S2=\exp(-\beta\|C1(j,k)-P_i\|^2), \tag{3}$$

where β is the sharpness of the exponential function, $C1(j,k)$ denotes the afferent C1 layer with scale j and orientation k, and P_i is a sampled patch from the previous C1 layers.

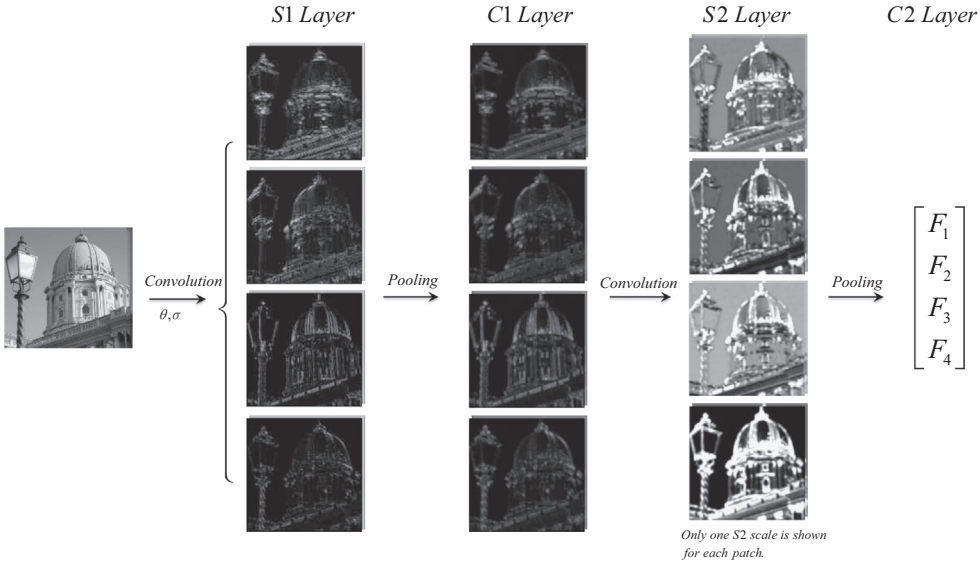

Fig. 1. HMAX structure overview.