高技术通讯 2014 年 第24 卷 第4 期:414 ~419

doi:10.3772/j.issn.1002-0470.2014.04.011

基于层次特征映射模型的目标识别①

余 鹏② 万里红 霍 宏 方 涛③

(上海交通大学自动化系系统控制与信息处理教育部重点实验室 上海 200240)

摘 要 为了较好地模拟生物视觉系统对复杂场景目标的感知特性,以提高目标识别水平,提出了一种新的受生物视觉信息处理的基本机理启发的前馈深度层次计算机视觉模型,即层次特征映射(HFM)模型。该模型利用高斯差分函数以及 Gabor 函数模拟初级视觉皮层中的方向图,并且采用竞争学习策略来学习更高层次神经元的感受野。实验表明,该模型可以很好地提取目标的特征和保留图像的主要信息,并且具有自学习的能力,能够在主流数据库上取得较好的识别结果,具有较好的发展前景。

关键词 目标识别,深度网络,方向图,竞争学习

0 引 言

近年来深度网络(用当前一层输出无监督训练获得新的一层[1]的思想发展而来)得到了大量研究。目前,有两类训练深度网络的方法,得到了普遍关注[2]:基于受限波尔兹曼机(restricted boltzmann machine,RBM)的方法[3,4]和基于卷积神经网络(convolutional neural networks,CNN)的方法[5-7]。其中,基于 CNN 的方法受到了视觉神经元感受野特性的启发,例如,在视觉皮层中,每个神经元都具有有限大小的感受野,而 CNN 则采用节点间稀疏连接的方式来模拟这一特性。为了进一步降低网络的训练复杂度,CNN 还采用了权值共享的方法。CNN 主要有卷积与子采样两种操作。通过这两种操作交替进行的方式,CNN 可获得对旋转、平移等的不变性,因而在图像、视频处理等领域获得了巨大成功。

然而,训练深度网络的算法多数集中在利用优化某一种能量函数来分层训练整个网络上。当前,各种深度网络模型更多地是从优化角度而不是从视皮层神经计算机理出发构造,从而导致建立的计算机视觉模型很难较好地模拟生物视觉系统对复杂场景目标的感知特性,影响了对目标的识别。尽管造成这一现象的原因有很多,但是层次最大化(Hierarchical MAX,HMAX)模型及其变种[8-10]既很好地切合了生物视觉的机理,并且取得了不错的效果。这说明,从神经计算的机理出发构建深度网络是可行的。基于这一认识,本文提出了一个受生物视觉启发的前馈深度层次模型,即层次特征映射(Hierarchical feature map,HFM)模型。该模型采用稀疏连接方式,分层学习,并且利用了方向图的最新研究成果。实验结果显示,模型可以很好地得到目标的抽象表示,能够取得较好的目标分类、识别结果。

1 HFM 模型框架

HFM 模型共有 4 层,即 S 层及其以上的 C1、C2 和 C3 层,其结构如图 1 所示,该模型利用系数连接来模拟神经元感受野。S 层中每个节点的感受野是由高斯差分函数或者 Gabor 函数来模拟的,同时,以特别的方式组合这些节点来模拟初级视觉皮层中的方向图。其余的 C1、C2 和 C3 三层节点的感受野特性通过竞争学习获得。层次愈高,每一层的节点感受野愈大。模型还利用稀疏连接来模拟神经元感受野。层次愈高,每一层的节点感受野愈大。

1.1 S 层:利用高斯差分函数和 Gabor 函数模拟方向图

S 层中每一个节点都是一个高斯差分函数或者

① 973 计划(2012CB719903),国家自然科学基金委创新研究群体(X198144),国家自然科学基金青年科学基金(41101386)和国家自然科学基金(41071256)资助项目。

② 男,1990 年生,硕士;研究方向:深度网络,计算机视觉;E-mail:adrianandyu@gmail.com

③ 通讯作者,E-mail:tfang@sjtu.edu.cn

(收稿日期:2013-09-15)

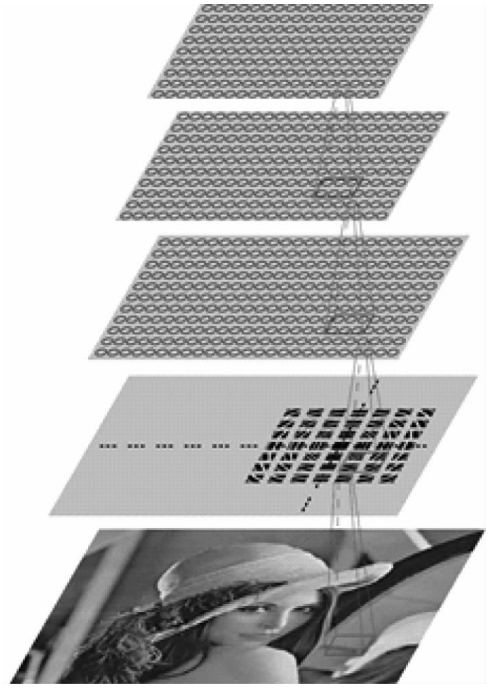


图 1　模型框架示意图

S 层中每个节点为高斯差分函数或者 Gabor 函数,C1、C2 和 C3 层节点为径向基函数(RBF),由竞争学习获得,随着层次的上升,高层次节点的感受野大小大约为低层次节点的感受野大小的 1.7 倍(图中红色方框只起示意作用,并不代表实际感受野大小)

Gabor 函数,并且本文还按照一定的拓扑结构组织这些节点的位置。文献[11]指出,初级视觉皮层 V1 层中的方向图可能是由 ON 和 OFF 中心型神经节细胞的位置分布决定的。图 2(a)展示出了基于这一思想可能存在的方向图,图 2(b)给出了对应的本文模型所采用的节点部分分布结构。在这样的一个结构中心节点为高斯差分函数,其余节点均为具有不同感受野特性的 Gabor 函数。每个节点感受野大小由实验决定,相邻节点的感受野互相重叠,步距为 1 个像素。整个 S 层就是由这样的结构重复组成的。高斯差分函数为

$$G(x,y) = \frac{1}{2\sigma_1^2}\exp\left(-\frac{x^2+y^2}{2\sigma_1^2}\right) - \frac{1}{2\sigma_2^2}\exp\left(-\frac{x^2+y^2}{2\sigma_2^2}\right) \tag{1}$$

式中,$\sigma_1 = \frac{1}{\sqrt{2}}, \sigma_2 = 3$。Gabor 函数为

$$G(x,y) = \exp\left(-\frac{X^2+\gamma^2Y^2}{2\sigma^2}\right)\times\sin\left(\frac{2\pi}{\lambda}X\right), s.t. \tag{2}$$

$$X = x\cos\theta + y\sin\theta, \quad Y = x\sin\theta - y\cos\theta \tag{3}$$

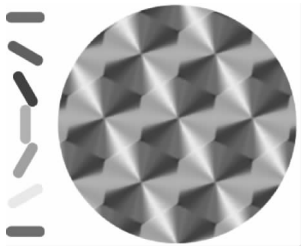
式中,$\gamma = 0.5, \sigma = 0.6\lambda$。在这一结构中,每一圈各条边中心处 Gabor 节点的 θ 分别为 $0, \frac{\pi}{2}, \pi, \frac{3\pi}{2}$。在同一圈内,$\theta$ 沿逆时针线性增加。由内而外,λ 也线性增加,具体的值由实验决定。并且,每一个高斯差分函数和 Gabor 函数由下式进行归一化:

$$P(x,y) = \frac{G(x,y)}{\|G\|_F} \tag{4}$$

式中,$\|\cdot\|_F$ 表示矩阵的 Frobenius 范数。对于任意一幅灰度图像(线性归一化至[0,1]区间),每个 S 层节点的响应 r 为

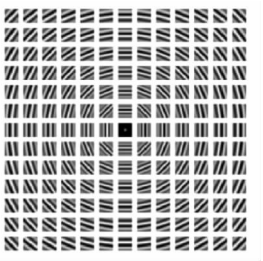
$$r = \sum_x\sum_y P(x,y)\cdot I_P(x,y) \tag{5}$$

式中,I_P 为每一个节点感受野内的图像块。



图中用不同的伪彩色表示不同的方向选择特性

(a)



一个高斯差分函数在结构的中心处,其余均为 Gabor 函数,这一结构主要用来模拟初级视觉皮层中的方向图

(b)

图 2　基于文献[11]思想的一种可能存在的方向图(a)和本文所使用的部分节点分布结构(b)

1.2　C1、C2 和 C3 层:更高层次视觉皮层

C1、C2 和 C3 层用来模拟更高层次视觉皮层。在高层次视觉皮层中,神经元的感受野大小一般比低层次神经元的感受野大一定的倍数[12]。在 HFM 模型中,高层次节点的感受野大小大约为低层次感受野大小的 1.7 倍。因此,随着层次的上升,每一层的规模显著减小。然而,对较大的输入图像来说,网络的规模依然可能变得很大。这时可以采用亚采样的技术:相邻节点的感受野相互重叠,但是步距为 2(或更多)而不为 1。这样的一个亚采样技术可以显著地降低模型的空间复杂度和计算复杂度,同时并

不会很大地削弱模型的性能。

C1、C2 和 C3 层节点的感受野为 RBF 函数[13]，其具体的值由竞争学习获得。每一个节点的响应 r 通过下式计算：

$$r = \exp(-\|X - W\|_F/\gamma) \quad (6)$$

式中，X 为底层节点的输出，W 为每个节点保存的模板（感受野特性），γ 控制节点的选择性强弱。

1.3 竞争学习

C1、C2 和 C3 层节点的感受野均由竞争学习获得。和一般的深度网络一样，网络是分层学习的。首先随机初始化当前训练层的所有节点。对每一个输入训练样本，模型计算出当前训练层的输出。然后训练层输出的所有局部极大值以及它们的 8 邻域位置被标记出来。对这些被标记出来的节点利用下式进行学习：

$$\Delta W = \alpha(X - W) \quad (7)$$

式中，学习速率 α 为一固定值 0.008。

需要说明的是，本文没有采用学习算法学习 S 层的节点，原因主要有以下几点：(1) 自然图像中的变化大，可能并不存在稳定的局部特征；(2) S 层实际模拟的是视觉信息从视网膜到 V1 层的处理过程，单一的算法可能不合适；(3) 文献[14]指出高层次视觉皮层（如 V4，IT）的可塑性比低层次视觉皮层（如 V1）强，学习主要发生在高层次视觉皮层。

2 实验及分析

为了验证本文提出的模型有效性，主要进行了两个实验。第一个实验通过图像重构，说明模型可以保留图像的主要结构信息。第二个实验给出模型在 Caltech-101 数据库[15]上的分类结果。

2.1 基于 S 层输出的图像重构实验

S 层的操作是线性的，所以总是可以找到一个矩阵 W（同时也是一个稀疏矩阵）来表示这样的一个线性变换。输入一幅图像 X（改写成向量的形式），其对应的 S 层输出 Y 可以写成

$$Y = WX \quad (8)$$

那么这一重构实验可以表达为：给定某一输入图像 X 对应的 S 层输出 Y，期望可以得到一个对 X 的重构 Z，使其满足

$$\min_z \frac{1}{2}\|Y - WZ\|_2^2 \quad s.t.\ 0 \leqslant Z \leqslant 1 \quad (9)$$

图 3 给出了两个重构实验的结果。为了降低计算复杂度，两个实验中的图像均被重新缩放到了 150×150 个像素。S 层中所有节点感受野大小均为 7×7 个像素，一个完整的如图 2(b) 所示的结构共有 13×13 个节点，Gabor 函数中 λ 的值由内而外从 1.5 线性增加至 4.5。从图 3 中可以看出，S 层的输出有效地保存了图像的主要内容，如线、边以及纹理等。为了定量地分析这一重构过程，本文采用结构相似度（structural SIMilarity，SSIM）指数[16]来进行评估。SSIM 指数比传统的峰值信噪比（PSNR）以及均方误差（mean square error，MSE）要更加贴合人类的视觉感知。为了获得两幅图像的相似程度，一般会在局部窗口计算 SSIM 指数，再求取其平均值（mean），即平均结构相似度（MSSIM）指数。可以计算出本文两个重构实验的 MSSIM 指数分别为 0.8066 和 0.9033。这说明本文的方法可以很好地保存图像信息。

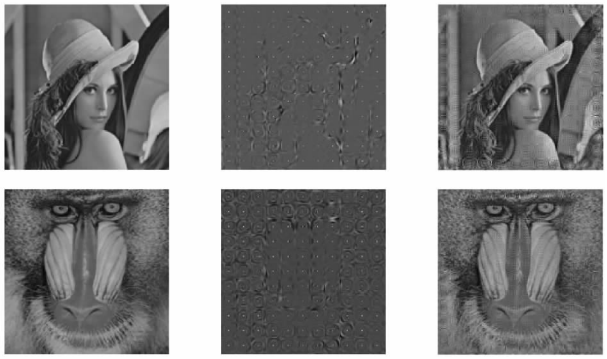


图 3 两个重构示例

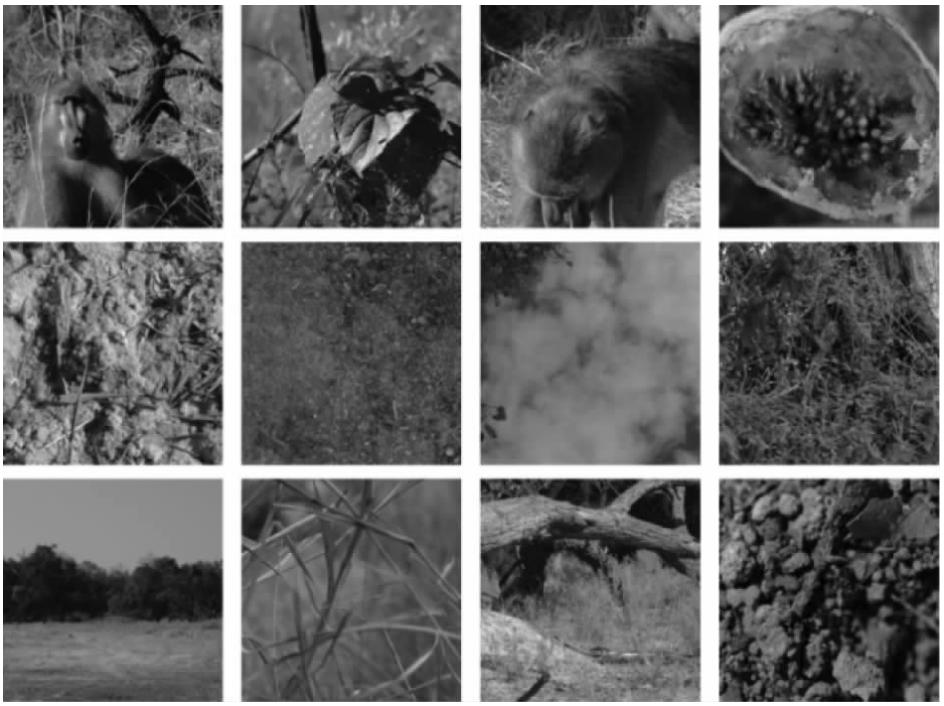
在每一行中，从左至右三幅图像依次为原始图像、对应的 S 层输出和基于重构算法得到的重构结果

2.2 Caltech-101 分类实验

为了展示模型具有自学习（self-taught learning）[17]的能力，模型利用 UPenn 自然图像库[18]进行训练。UPenn 自然图像库摄制于与人眼进化环境相似的地点，包含场景复杂多样。本文将这一数据库中对比度很低的图像舍弃，余下 5172 幅图像作为模型训练集。而 Caltech-101 数据集的所有图像均被使用了，包括 102 个类，共 9144 幅图像。

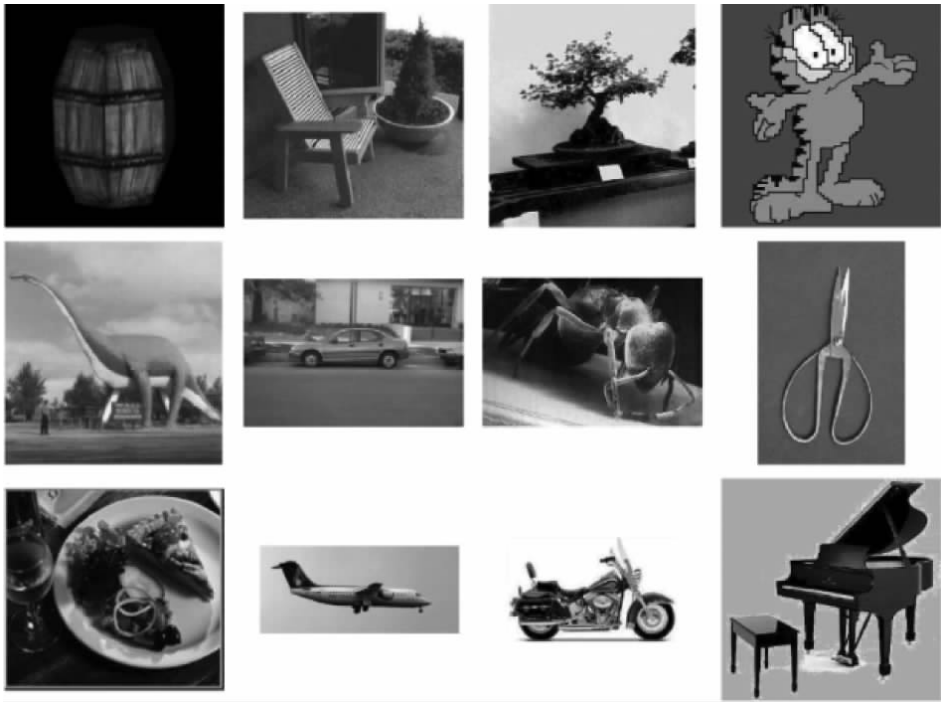
所有的图像在使用前均被转换成灰度图像。由于 UPenn 自然图像库中图像大小均为 3008×2000 个像素，而 Caltech-101 数据库中图像大小约在 300×300 个像素附近，所以模型的输入设置为 300×300 个像素。UPenn 自然图像库中的图像首先被缩放至原先的 1/4，再提取图像中心处的 300×300 个像素。Caltech-101 数据库中图像较长的边首

先被放大或缩小至 300 个像素,并保持图像原先的从横比,再将整幅图像放置到 300×300 个像素大小的白色背景上。图 4(a)展示了部分实际使用的 UPenn 库中的图像,图 4(b)展示了部分实际使用的 Caltech-101 库中的图像。



所有图像均转化为灰度图像,再经过缩放处理,最后提取中心部分

(a)UPenn 自然图像库图像



所有图像均转化为灰度图像,再经过放大或缩小处理,最后放至白色背景中心处

(b)Caltech-101 图像库图像

图 4　UPenn 自然图像库图像和 Caltech-101 图像库图像

2.2.1 C 层感受野

为了达到较好的收敛结果,模型中每层所需的学习样本数大约为该层节点数的 30 倍左右。这样,在单片 Intel Xeon X5690 CPU 上利用 Intel TBB 进行加速运算,整个训练过程大约需要 4.6h,完整处理一幅图像所需的平均时间大约为 42ms(不包括文件读写时间)。为了得到 C 层各节点的感受野,本文利用计算底层感受野加权和的方式得到高层感受野的形状。图 5 给出了在 UPenn 自然图像库上学习得到的 C 层感受野。从图中可以看出:C1 层感受野多为转折点的形状;C2 层感受野形状复杂,为多个转折点的汇集;C3 层感受野的形状为不同 C2 层节点的组合形式。每一层相邻节点的感受野具有一定的相似性,与大脑皮层内的功能图类似。

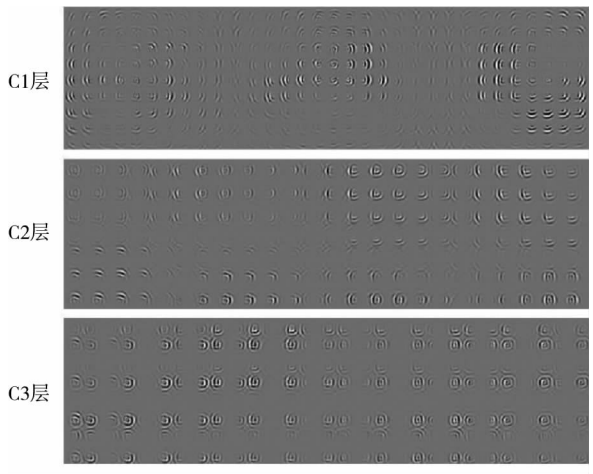


图 5 C1、C2 和 C3 层感受野形状

2.2.2 分类实验

模型首先利用 UPenn 数据库进行训练,再在 Caltech-101 数据库中利用 RBF SVM[19] 进行分类实验,利用 5 折交叉检验确定 SVM 各参数的值。分类实验的流程和文献[4]一样:首先从 Caltech-101 数据库的每一类中随机选择 30 幅训练样本和 30 幅测试样本,再给出平均的分类精度,即被正确分类的样本数与全部样本数之比,一共重复 10 次实验。表 1 给出了模型的最终精度,以及 HMAX 模型和一些深度网络在 Caltech-101 上的精度。

从表 1 可以看出,本文的模型显著好于 HMAX 模型,这主要是由于 HMAX 模型没有采用有效的学习方法来学习所用的滤波器。本文的模型和一般的深度网络相当,说明所采用的策略是有效的。然而需要指出的是深度网络已经得到了长足地发展,本文的模型距离先进的深度网络还有着很大的差距。但是鉴于 HFM 模型还有很大的发展潜力,未来的研究还是很有必要的。

表 1 Caltech-101 上的分类精度

方法	分类精度
文献[8]	42%
文献[9]	56%
文献[4]	65.4%
文献[20]	65.6%
文献[21]	71.0%
HFM (本文)	69.54%

3 结 论

本文受生物视觉信息处理的基本机理启发,结合最新的研究成果,利用竞争学习策略建立了一个生物学上可行的深度层次模型——HFM 模型。由于采用了稀疏连接以及亚采样技术,模型本身的计算复杂度较低。实验显示,本文提出的 HFM 模型能够很好地提取目标的抽象特征,并取得了较好地效果。未来的主要工作应该集中在如何有效地将反馈连接加入模型之中。

参考文献

[1] Bengio Y, Courville A, Vincent, P. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2013, 35(8): 1798-1828

[2] Arel I, Rose D C, Karnowski T P. Deep machine learning-A new frontier in artificial intelligence research. *IEEE Computational Intelligence Magazine*, 2010, 5(4): 13-18

[3] Hinton G E, Osindero S, Teh Y W. A fast learning algorithm for deep belief nets. *Neural computation*, 2006, 18(7): 1527-1554

[4] Lee H, Grosse R, Ranganath R, et al. Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations. In: Proceedings of the 26th Annual International Conference on Machine Learning, Montreal, Canada, 2009. 609-616

[5] LeCun Y, Bottou L, Bengio Y, et al. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 1998, 86(11): 2278-2324

[6] Le Q V, Ngiam J, Chen Z, et al. Tiled convolutional neural

networks. In: Advances in Neural Information Processing Systems, Vancouver, Canada, 2010. 1279-1287

[7] 许可. 卷积神经网络在图像识别上的应用的研究:[硕士学位论文]. 杭州:浙江大学计算机科学与技术学院,2012. 3-26

[8] Serre T, Wolf L, Bileschi S, et al. Robust object recognition with cortex-like mechanisms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2007, 29(3): 411-426

[9] Mutch J, Lowe D G. Object class recognition and localization using sparse features with limited receptive fields. *International Journal of Computer Vision*, 2008, 80(1): 45-57

[10] 江达秀. 基于 HMAX 模型的人脸表情识别研究:[硕士学位论文]. 杭州:浙江理工大学信息电子学院,2010. 34-42

[11] Paik S B, Ringach D L. Link between orientation and retinotopic maps in primary visual cortex. *Proceedings of the National Academy of Sciences*, 2012, 109(18): 7091-7096

[12] Freeman J, Simoncelli E P. Metamers of the ventral stream. *Nature neuroscience*, 2011, 14(9): 1195-1201

[13] Poggio T, Bizzi E. Generalization in vision and motor control. *Nature*, 2004, 431(7010): 768-774

[14] Li N, DiCarlo J J. Unsupervised natural experience rapidly alters invariant object representation in visual cortex. *Science*, 2008, 321(5895): 1502-1507

[15] Li F F, FergusR, PeronaP. Learning generativevisual models from few training examples: an incrementalBayesian approach tested on 101 object categories. In: Proceedings of the IEEECVPR Workshop on Generative-Model Based Vision, Washington, USA, 2004.

[16] Wang Z, Bovik A C, SheikhH R, et al. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 2004, 13(4): 600-612

[17] Raina R, Battle A, Lee H, et al. Self-taught learning: transfer learning from unlabeled data. In: Proceedings of the 24th International Conference on Machine Learning, Corvallis, USA, 2007. 759-766

[18] Tkacik G, Garrigan P, Ratliff C, et al. Natural images from the birthplace of the human eye. *PLoS One*, 2011, 6(6): e20409. doi: 10. 1371/journal. pone. 0020409

[19] Chang C C, Lin C J. LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2011, 2(3): 27: 1-27: 27

[20] Jarrett K, Kavukcuoglu K, Ranzato M A, et al. What is the best multi-stage architecture for object recognition?. In: Proceedings of the 12th International Conference on Computer Vision, Kyoto, Japan, 2009. 2146-2153

[21] Zeiler M D, Taylor G W, FergusR. Adaptive deconvolutional networks for mid and high level feature learning. In: Proceedings of the 13th International Conference on Computer Vision, Barcelona, Spain, 2011. 2018-2025

Object recognition based on a hierarchical feature map model

Yu Peng, Wan Lihong, Huo Hong, Fang Tao

(Key Laboratory of System Control and Information Processing, Ministry of Education, Department of Automation, Shanghai Jiao Tong University, Shanghai 200240)

Abstract

A new biologically inspired feed-forward deep hierarchical model, i. e. the hierarchical feature map (HFM) model, is introduced to better simulate the biological vision system's perception of objects in a complex scene for improvement of the visual object recognition. The HFM model uses the Difference of Gaussian function and Gabor function to simulate the orientation map in the primary visual cortex V1, and adopts a competitive learning strategy to learn the receptive field (RF) properties of higher level neurons. The experimental results show that the HFM model could well preserve the main structure of images. The model is also capable of self-taught learning and can achieve promising results on popular image databases, showing a good prospect for development.

Key words: object recognition, deep network, orientation map, competitive learning