图像物体分类与检测算法综述

黄凯奇 任伟强 谭铁牛

（中国科学院自动化研究所模式识别国家重点实验室智能感知与计算研究中心 北京 100190）

摘 要 图像物体分类与检测是计算机视觉研究中的两个重要的基本问题，也是图像分割、物体跟踪、行为分析等其他高层视觉任务的基础. 该文从物体分类与检测问题的基本定义出发，首先从实例、类别、语义三个层次对物体分类与检测研究中存在的困难与挑战进行了阐述. 接下来，该文以物体检测和分类方面的典型数据库和国际视觉算法竞赛 PASCAL VOC 竞赛为主线对近年来物体分类与检测的发展脉络进行了梳理与总结，指出表达学习和结构学习在于物体分类与检测中占有重要的地位. 最后文中对物体分类与检测的发展方向进行了思考和讨论，探讨了图像物体识别中下一步研究可能的方向.

关键词 物体分类；物体检测；计算机视觉；特征表达；结构学习

中图法分类号 TP391 **DOI 号** 10.3724/SP.J.1016.2014.01225

A Review on Image Object Classification and Detection

HUANG Kai-Qi REN Wei-Qiang TAN Tie-Niu

(*Center for Research on Intelligent Perception and Computing, National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing* 100190)

Abstract Image object classification and detection are two of the most essential problems in computer vision. They are the basis of many other complex vision problems, such as segmentation, tracking, and action analysis. In this paper, we try to give an analysis of object classification and detection algorithms based on PASCAL VOC challenge, which is generally acknowledged as a public evaluation for object recognition. We first discuss the importance of object classification and detection; next we summarize the difficulties and challenges in the development of basic object recognition. Then we review the yearly achievements in the study of object classification and detection. Finally we discuss the development directions of object classification and detection, from the view of representations learning and structure learning.

Keywords object classification; object detection; computer vision; feature representations; structural learning

1 图像物体分类与检测概述

物体分类与检测是计算机视觉、模式识别与机器学习领域非常活跃的研究方向. 物体分类与检测在很多领域有广泛应用，包括安防领域的人脸识别、行人检测、智能视频分析、行人跟踪等，交通领域的交通场景物体识别、车辆计数、逆行检测、车牌检测

收稿日期：2013-09-08；最终修改稿收到日期：2013-12-25. 本课题得到国家“九七三”重点基础研究发展规划项目基金（2012CB316302）、国家自然科学基金（61322209）和国家科技支撑计划（2012BAH07B01）资助. **黄凯奇**，男，1977 年生，博士，研究员，国家自然科学基金优秀青年基金获得者、中国计算机学会（CCF）高级会员，曾担任 IEEE 北京分会副秘书长，主要研究领域为计算机视觉、模式识别、视觉监控. E-mail：kqhuang@nlpr.ia.ac.cn. **任伟强**，男，1985 年生，博士研究生，主要研究方向为计算机视觉、模式识别. **谭铁牛**，男，1964 年生，博士，研究员，中国科学院院士，主要研究领域为生物特征识别、智能视频监控、网络数据理解与安全.

与识别，以及互联网领域的基于内容的图像检索、相册自动归类等. 可以说，物体分类与检测已经应用于人们日常生活的方方面面，计算机自动分类与检测技术也在一定程度了减轻了人的负担，改变了人类的生活方式.

计算机视觉理论的奠基者，英国神经生理学家Marr[1]认为，视觉要解决的问题可归结为“What is Where”，即“什么东西在什么地方”. 因此计算机视觉的研究中，物体分类和检测是最基本的研究问题之一. 如图 1 所示，给定一张图片，物体分类要回答的问题是这张图片中是否包含某类物体(比如牛)；物体检测要回答的问题则是物体出现在图中的什么地方，即需要给出物体的外接矩形框，如图 1(b)所示. 物体分类与检测的研究，是整个计算机视觉研究的基石，是解决跟踪、分割、场景理解等其他复杂视觉问题的基础. 欲对实际复杂场景进行自动分析与理解，首先就需要确定图像中存在什么物体(分类问题)，或者是确定图像中什么位置存在什么物体(检测问题). 鉴于物体分类与检测在计算机视觉领域的重要地位，研究鲁棒、准确的物体分类与检测算法，无疑有着重要的理论意义和实际意义.

(a) 物体分类 (b) 物体检测

图 1 视觉识别中的物体分类与检测

本文从物体分类与检测问题的基本定义出发，首先从实例、类别、语义三个层次对物体分类与检测研究中存在的困难与挑战进行了阐述. 接下来，本文以物体检测和分类方面的主流数据库和国际视觉算法竞赛 PASCAL VOC 竞赛为主线对近年来物体分类与检测算法的发展脉络进行了梳理与总结，总结了物体分类与检测算法的主流方法：基于表达学习和结构学习. 在此基础上，本文对物体分类与检测算法的发展方向进行了思考和讨论，指出了物体检测和物体分类算法的有机统一，探讨了下一步研究的方向.

2 物体分类与检测的难点与挑战

物体分类与检测是视觉研究中的基本问题，也是一个非常具有挑战性的问题. 物体分类与检测的难点与挑战在本文中分为 3 个层次：实例层次、类别层次和语义层次，如图 2 所示.

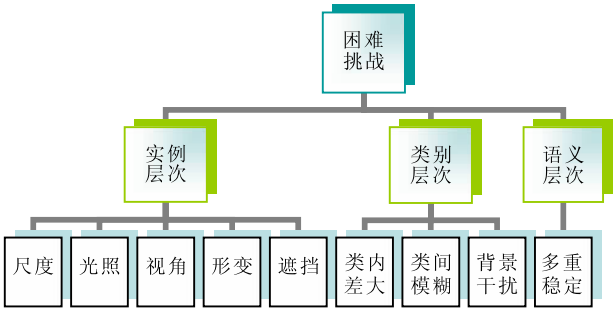

图 2 物体分类与检测研究存在的困难与挑战

(1) 实例层次. 针对单个物体实例而言，通常由于图像采集过程中光照条件、拍摄视角、距离的不同、物体自身的非刚体形变以及其他物体的部分遮挡，使得物体实例的表观特征产生很大的变化，给视觉识别算法带来了极大的困难.

(2) 类别层次. 困难与挑战通常来自 3 个方面，首先是类内差大，也即属于同一类的物体表观特征差别比较大，其原因有前面提到的各种实例层次的变化，但这里更强调的是类内不同实例的差别，例如图 3(a)所示，同样是椅子，外观却是千差万别，而从语义上来讲，具有“坐”的功能的器具都可以称为椅子；其次是类间模糊性，即不同类的物体实例具有一定的相似性，如图 3(b)所示，左边的是一只狼，右边的是一只哈士奇，但我们从外观上却很难分开二者；再次是背景的干扰，在实际场景下，物体不可能出现在一个非常干净的背景下，往往相反，背景可能是非常复杂的、对我们感兴趣的物体存在干扰的，这使得识别问题的难度大大增加.

(3) 语义层次. 困难和挑战与图像的视觉语义相关，这个层次的困难往往非常难处理，特别是对现在的计算机视觉理论水平而言，一个典型的问题称为多重稳定性. 如图 3 所示，图 3(c)左边既可以看

(a) (b) (c)

图 3 分类与检测存在挑战的例子

成是两个面对面的人，也可以看成是一个燃烧的蜡烛；右边则同时可以解释为兔子或者小鸭. 同样的图像，不同的解释，这既与人的观察视角、关注点等物理条件有关，也与人的性格、经历等有关，而这恰恰是视觉识别系统难以处理的部分.

3　物体分类与检测数据库

数据是视觉识别研究中最重要的因素之一，通常我们更多关注于模型、算法本身，事实上，数据在视觉任务中的作用越来越明显. 大数据时代的到来，也使得研究人员开始更加重视数据. 在数据足够多的情况下，我们甚至可以使用最简单的模型、算法，比如最近邻分类、朴素贝叶斯分类器都能得到很好的效果. 鉴于数据对算法的重要性，我们将在本节对视觉研究中物体分类与检测方面的主流数据进行概述，从中也可以一窥目标分类、检测的发展. 在介绍不同数据库时，将主要从数据库图像数目、类别数目、每类样本数目、图像大小、分类检测任务难度等方面进行阐述，如表 1 所示.

表 1　主流物体分类与识别数据库

数据库	图像数目	类别数目	每类样本数目	图像大小(pixel)	难度
MNIST[68]	60000	10	6000	28×28	容易
CIFAR-10[5]	60000	10	6000	32×32	中等
MPEG-7[4]	1400	70	20	256×256～650×600	中等
15 Scenes[9]	4485	15	200～400	约 300×250	容易
Caltech-101[7]	9146	101	40～800	约 300×200	中等
Caltech-256[8]	30607	256	80+	约 300×200	较难
PASCAL VOC 2007[11]	9963	20	96～2008	约 470×380	很难
SUN397[13]	108754	397	100+	约 500×300	很难
SUN2012[13]	16873	8	2000	约 500×300	很难
Tiny Images[6]	7900 万	75062	—	32×32	很难
ImageNet-1000[12]	120 万	1000	—	约 500×400	较难
ImageNet[12]	1400 万	10 万	1000	约 500×400	很难

早期物体分类研究集中于一些较为简单的特定任务，如 OCR、形状分类等. OCR 中数字手写识别是一个得到广泛研究的课题，相关数据库中最著名的是 MNIST[2] 数据库. MNIST 是一个数字手写识别领域的标准评测数据集，数据库大小是 60000，一共包含 10 类阿拉伯数字，每类提供 5000 张图像进行训练，1000 张进行测试. MNIST 的图像大小为 28×28，即 784 维，所有图像为手写数字，存在较大的形变. 形状分类是另一个比较重要的物体分类初期的研究领域，相关数据库有 ETHZ Shape Classes[3]、MPEG-7[4] 等. 其中 ETHZ Shape Classes 包含 6 类具有较大差别的形状类别：苹果、商标、瓶子、长颈鹿、杯子、天鹅，整个数据库包含 255 张测试图像.

CIFAR-10[3] 和 CIFAR-100[5] 数据库是 Tiny images[6] 的两个子集，分别包含了 10 类和 100 类物体类别. 这两个数据库的图像尺寸都是 32×32，而且是彩色图像. CIFAR-10 包含 6 万的图像，其中 5 万用于模型训练，1 万用于测试，每一类物体有 5000 张图像用于训练，1000 张图像用于测试. CIFAR-100 与 CIFAR-10 组成类似，不同的是包含了更多的类别：20 个大类，大类又细分为 100 个小类别，每类包含 600 张图像. CIFAR-10 和 CIFAR-100 数据库尺寸较小，但是数据规模相对较大，非常适合复杂模型特别是深度学习模型训练，因而成为深度学习领域主流的物体识别评测数据集.

Caltech-101[7] 是第一个规模较大的一般物体识别标准数据库，除背景类别外，它一共包含了 101 类物体，共 9146 张图像，每类中图像数目从 40 到 800 不等，图像尺寸也达到 300 左右. Caltech-101 是以物体为中心构建的数据库，每张图像基本只包含一个物体实例，且居于图像中间位置. 物体尺寸相对图像尺寸比例较大，且变化相对实际场景来说不大，比较容易识别. Caltech-101 每类的图像数目差别较大，有些类别只有很少的训练图像，也约束了可以使用的训练集大小. Caltech 256[8] 与 Caltech-101 类似，区别是物体类别从 101 类增加到了 256 类，每类包含至少 80 张图像. 图像类别的增加，也使得 Caltech-256 上的识别任务更加困难，使其成为检验算法性能与扩展性的新基准. 15 Scenes 是由 Lazebnik 等人[9] 在 Li 等人[10] 的 13 Scenes 数据库的基础上加入了两个新的场景构成的，一共有 15 个自然场景，4485 张图像，每类大概包含 200～400 张图像，图像分辨率约为 300×250. 15 Scenes 数据库主要用于场景分类评测，由于物体分类与场景分类在模型与算法上差别不大，该数据库也在图像分类问题上得到广泛的使用.

PASCAL VOC[11]从 2005 年到 2012 年每年都发布关于分类、检测、分割等任务的数据库，并在相应数据库上举行了算法竞赛，极大地推动了视觉研究的发展进步. 最初 2005 年 PASCAL VOC 数据库只包含人、自行车、摩托车、汽车共 4 类，2006 年类别数目增加到 10 类，2007 年开始类别数目固定为 20 类，以后每年只增加部分样本. PASCAL VOC 数据库中物体类别均为日常生活中常见的物体，如交通工具、室内家具、人、动物等. PASCAL VOC 2007 数据库共包含 9963 张图片，图片来源包括 Filker 等互联网站点以及其他数据库，每类大概包含 96～2008 张图像，均为一般尺寸的自然图像. PASCAL VOC 数据库与 Caltech-101 相比，虽然类别数更少，但由于图像中物体变化极大，每张图像可能包含多个不同类别物体实例，且物体尺度变化很大，因而分类与检测难度都非常大. 该数据库的提出，对物体分类与检测的算法提出了极大的挑战，也催生了大批优秀的理论与算法，将物体识别的研究推向了一个新的高度.

随着分类与检测算法的进步，很多算法在以上提到的相关数据库上性能都接近饱和，同时随着大数据时代的到来、硬件技术的发展，也使得在更大规模的数据库上进行研究和评测成为必然. ImageNet[12]是由 Li 主持构建的大规模图像数据库，图像类别按照 WordNet 构建，全库截至 2013 年共有 1400 万张图像，2.2 万个类别，平均每类包含 1000 张图像. 这是目前视觉识别领域最大的有标注的自然图像分辨率的数据集，尽管图像本身基本还是以目标为中心构建的，但是海量的数据和海量的图像类别，使得该数据库上的分类任务依然极具挑战性. 除此之外，ImageNet 还构建了一个包含 1000 类物体 120 万图像的子集，并以此作为 ImageNet 大尺度视觉识别竞赛的数据平台，也逐渐成为物体分类算法评测的标准数据集.

SUN 数据库[13]的构建是希望给研究人员提供一个覆盖较大场景、位置、人物变化的数据库，库中的场景名是从 WordNet 中的所有场景名称中得来的. SUN 数据库包含两个评测集，一个是场景识别数据集，称为 SUN-397，共包含 397 类场景，每类至少包含 100 张图片，总共有 108 754 张图像. 另一个评测集为物体检测数据集，称为 SUN2012，包含 16 873 张图像.

Tiny images[6]是一个图像规模更大的数据库，共包含 7900 万张 32×32 图像，图像类别数目有 7.5 万，尽管图像分辨率较低，但还是具有较高的区分度，而其绝无仅有的数据规模，使其成为大规模分类、检索算法的研究基础.

我们通过分析表 1 可以看到，在物体分类的发展过程中，数据库的构建大致可以分为 3 个阶段，经历了一个从简单到复杂，从特殊到一般，从小规模到大规模的跨越. 早期的手写数字识别 MNIST、形状分类 MPEG-7 等都是研究特定问题中图像分类，之后研究人员开始进行更广泛的一般目标分类与检测的研究，典型的数据库包括 15 Scenes、Caltech-101/256、PASCAL VOC 2007 等；随着词包模型等算法的发展与成熟，更大规模的物体分类与检测研究得到了广泛的关注，这一阶段的典型数据库包括 SUN 数据库、ImageNet 以及 Tiny 等. 近年来，数据库构建中的科学性也受到越来越多的关注，Torralba 等人[14]对数据库的 Bias、泛化性能、价值等问题进行了深入的讨论，提出排除数据库构建过程中的选择偏好、拍摄偏好、负样本集偏好是构造更加接近真实视觉世界的视觉数据库中的关键问题. 伴随着视觉处理理论的进步，视觉识别逐渐开始处理更加真实场景的视觉问题，因而对视觉数据库的泛化性、规模等也提出了新的要求和挑战.

我们也可以发现，物体类别越多，导致类间差越小，分类与检测任务越困难，图像数目、图像尺寸的大小，则直接对算法的可扩展性提出了更高的要求，如何在有限时间内高效地处理海量数据、进行准确的目标分类与检测成为当前研究的热点.

4 物体分类与检测的发展历程

图像物体识别的研究已经有五十多年的历史. 各类理论和算法层出不穷，在这部分，我们对物体分类与检测的发展脉络进行了简单梳理，并将其中里程碑式的工作进行综述. 特别的，我们以国际视觉算法竞赛 PASCAL VOC 竞赛[11]为主线对物体分类与检测算法近年来的主要进展进行综述，这个系列的竞赛对物体识别研究的发展影响深远，其工作也代表了当时的最高水平.

物体分类任务要求回答一张图像中是否包含某种物体，对图像进行特征描述是物体分类的主要研究内容. 一般说来，物体分类算法通过手工特征或者特征学习方法对整个图像进行全局描述，然后使用分类器判断是否存在某类物体. 物体检测任务则更为复杂，它需要回答一张图像中在什么位置存在一个什么物体，因而除特征表达外，物体结构是物体检

测任务不同于物体分类的最重要之处. 总的来说，近年来物体分类方法多侧重于学习特征表达，典型的包括词包模型(Bag-of-Words)、深度学习模型；物体检测方法则侧重于结构学习，以形变部件模型为代表. 这里我们首先以典型的分类检测模型来阐述其一般方法和过程，之后以 PASCAL VOC(包含ImageNet)竞赛历年来的最好成绩来介绍物体分类和物体检测算法的发展，包括物体分类中的词包模型、深度学习模型以及物体检测中的结构学习模型，并分别对各个部分进行阐述.

4.1　基于词包模型的物体分类

从表 2 我们可以发现，词包模型是 VOC 竞赛中物体分类算法的基本框架，几乎所有的参赛算法都是基于词包模型. 我们将从底层特征、特征编码、空间约束、分类器设计、模型融合几个方面来展开阐述.

表 2　历年 PASCAL VOC 竞赛分类算法

年份	底层特征	特征编码	空间约束	分类器	融合
2005	密集 SIFT	向量量化	无	线性 SVM	特征拼接
2006	兴趣点检测＋密集提取	向量量化	SPM	两层核 SVM	两层融合
2007	密集＋兴趣点，多特征	向量量化	SPM	核 SVM	多特征，通道加权
2008	密集＋兴趣点，多特征	软量化	SPM	多分类器	多特征，多分类器
2009	密集 SIFT	GMM，LCC	SPM	线性 SVM	多特征
2010	密集＋兴趣点，多特征	向量量化	SPM，检测	多分类器	多特征，多分类器，分割，检测
2011	密集＋兴趣点，多特征	向量量化	SPM，检测	多分类器	多特征，多分类器，分割，检测
2012	密集＋兴趣点，多特征	向量量化，Fisher 向量	SPM，检测	多分类器	多特征，多分类器，分割，检测

词包模型(Bag-of-Words)最初产生于自然语言处理领域，通过建模文档中单词出现的频率来对文档进行描述与表达. Csurka 等人[15]于 2004 年首次将词包的概念引入计算机视觉领域，由此开始大量的研究工作集中于词包模型的研究，并逐渐形成了由下面 4 部分组成的标准物体分类框架：

(1) 底层特征提取. 底层特征是物体分类与检测框架中的第一步，底层特征提取方式有两种：一种是基于兴趣点检测，另一种是采用密集提取的方式. 兴趣点检测算法通过某种准则选择具有明确定义的、局部纹理特征比较明显的像素点、边缘、角点、区块等，并且通常能够获得一定的几何不变性，从而可以在较小的开销下得到更有意义的表达，最常用的兴趣点检测算子有 Harris 角点检测子、FAST (Features from Accelerated Segment Test)算子、LoG(Laplacian of Gaussian)、DoG(Difference of Gaussian)等. 近年来物体分类领域使用更多的则是密集提取的方式，从图像中按固定的步长、尺度提取出大量的局部特征描述，大量的局部描述尽管具有更高的冗余度，但信息更加丰富，后面再使用词包模型进行有效表达后通常可以得到比兴趣点检测更好的性能. 常用的局部特征包括 SIFT(Scale-Invariant Feature Transform，尺度不变特征转换)[16]、HOG (Histogram of Oriented Gradient，方向梯度直方图)[17]、LBP(Local Binary Pattern，局部二值模式)[18]等. 从表 2 可以看出，历年最好的物体分类算法都采用了多种特征，采样方式上密集提取与兴趣点检测相结合，底层特征描述也采用了多种特征描述子，这样做的好处是，在底层特征提取阶段，通过提取到大量的冗余特征，最大限度的对图像进行底层描述，防止丢失过多的有用信息，这些底层描述中的冗余信息主要靠后面的特征编码和特征汇聚得到抽象和简并. 事实上，近年来得到广泛关注的深度学习理论中一个重要的观点就是手工设计的底层特征描述子作为视觉信息处理的第一步，往往会过早地丢失有用的信息，直接从图像像素学习到任务相关的特征描述是比手工特征更为有效的手段.

(2) 特征编码. 密集提取的底层特征中包含了大量的冗余与噪声，为提高特征表达的鲁棒性，需要使用一种特征变换算法对底层特征进行编码，从而获得更具区分性、更加鲁棒的特征表达，这一步对物体识别的性能具有至关重要的作用，因而大量的研究工作都集中在寻找更加强大的特征编码方法，重要的特征编码算法包括向量量化编码[19]、核词典编码[20]、稀疏编码[21]、局部线性约束编码[22]、显著性编码[23]、Fisher 向量编码[24]、超向量编码[25]等. 最简单的特征编码是向量量化编码[19]，它的出现甚至比词包模型的提出还要早. 向量量化编码是通过一种量化的思想，使用一个较小的特征集合(视觉词典)来对底层特征进行描述，达到特征压缩的目的. 向量量化编码只在最近的视觉单词上响应为 1，因而又称为硬量化编码、硬投票编码，这意味着向量量化编码只能对局部特征进行很粗糙的重构. 但向量量化编码思想简单、直观，也比较容易高效实现，因

而从 2005 年第一届 PASCAL VOC 竞赛以来，就得到了广泛的使用. 在实际图像中，图像局部特征常常存在一定的模糊性，即一个局部特征可能和多个视觉单词差别很小，这个时候若使用向量量化编码将只利用距离最近的视觉单词，而忽略了其他相似性很高的视觉单词. 为了克服这种模糊性问题，van Gemert 等人[20]提出了软量化编码(又称核视觉词典编码)算法，局部特征不再使用一个视觉单词描述，而是由距离最近的 K 个视觉单词加权后进行描述，有效解决了视觉单词的模糊性问题，提高了物体识别的精度. 稀疏表达理论[21]近年来在视觉研究领域得到了大量的关注，研究人员最初在生理实验中发现细胞在绝大部分时间内是处于不活动状态，也即在时间轴上细胞的激活信号是稀疏的. 稀疏编码通过最小二乘重构加入稀疏约束来实现在一个过完备基上响应的稀疏性. ℓ_0 约束是最直接的稀疏约束，但通常很难进行优化，近年来更多使用的是 ℓ_1 约束，可以更加有效地进行迭代优化，得到稀疏表达. 2009 年 Yang 等人[26]将稀疏编码应用到物体分类领域，替代了之前的向量量化编码和软量化编码，得到一个高维的高度稀疏的特征表达，大大提高了特征表达的线性可分性，仅仅使用线性分类器就得到了当时最好的物体分类结果，将物体分类的研究推向了一个新的高度上. 稀疏编码在物体分类上的成功也不难理解，对于一个很大的特征集合(视觉词典)，一个物体通常只和其中较少的特征有关，例如，自行车通常和表达车轮、车把等部分的视觉单词密切相关，与飞机机翼、电视机屏幕等关系很小，而行人则通常在头、四肢等对应的视觉单词上有强响应. 稀疏编码存在一个问题，即相似的局部特征可能经过稀疏编码后在不同的视觉单词上产生响应，这种变换的不连续性必然会产生编码后特征的不匹配，影响特征的区分性能. 局部线性约束编码[22]的提出就是为了解决这一问题，它通过加入局部线性约束，在一个局部流形上对底层特征进行编码重构，这样既可以保证得到的特征编码不会有稀疏编码存在的不连续问题，也保持了稀疏编码的特征稀疏性. 局部线性约束编码中，局部性是局部线性约束编码中的一个核心思想，通过引入局部性，一定程度上改善了特征编码过程的连续性问题，即距离相近的局部特征在经过编码之后应该依然能够落在一个局部流形上. 局部线性约束编码可以得到稀疏的特征表达，与稀疏编码不同之处就在于稀疏编码无法保证相近的局部特征编码之后落在相近的局部流形. 从表 2 可以看出，2009 年的分类竞赛冠军采用了混合高斯模型聚类和局部坐标编码(局部线性约束编码是其简化版本)，仅仅使用线性分类器就取得了非常好的性能. 不同于稀疏编码和局部线性约束编码，显著性编码[23]引入了视觉显著性的概念，如果一个局部特征到最近和次近的视觉单词的距离差别很小，则认为这个局部特征是不“显著的”，从而编码后的响应也很小. 显著性编码通过这样很简单的编码操作，在 Caltech101/256，PASCAL VOC 2007 等数据库上取得了非常好的结果，而且由于是解析的结果，编码速度也比稀疏编码快很多. Huang 等人[23]发现显著性表达配合最大值汇聚在特征编码中有重要的作用，并认为这正是稀疏编码、局部约束线性编码等之所以在图像分类任务上取得成功的原因. 超向量编码[25]，Fisher 向量编码[24]是近年提出的性能最好的特征编码方法，其基本思想有相似之处，都可以认为是编码局部特征和视觉单词的差. Fisher 向量编码同时融合了产生式模型和判别式模型的能力，与传统的基于重构的特征编码方法不同，它记录了局部特征与视觉单词之间的一阶差分和二阶差分. 超向量编码则直接使用局部特征与最近的视觉单词的差来替换之前简单的硬投票. 这种特征编码方式得到的特征向量表达通常是传统基于重构编码方法的 M 倍(M 是局部特征的维度). 尽管特征维度要高出很多，超向量编码和 Fisher 向量编码在 PASCAL VOC、ImageNet 等极具挑战性、大尺度数据库上获得了当时最好的性能，并在图像标注、图像分类、图像检索等领域得到应用. 2011 年 ImageNet 分类竞赛冠军采用了超向量编码，2012 年 VOC 竞赛冠军则是采用了向量量化编码和 Fisher 向量编码.

(3)特征汇聚. 空间特征汇聚是特征编码后进行的特征集整合操作，通过对编码后的特征，每一维都取其最大值或者平均值，得到一个紧致的特征向量作为图像的特征表达. 这一步得到的图像表达可以获得一定的特征不变性，同时也避免了使用 特征集进行图像表达的高额代价. 最大值汇聚在绝大部分情况下的性能要优于平均值汇聚，也在物体分类中使用最为广泛. 由于图像通常具有极强的空间结构约束，空间金字塔匹配(Spatial Pyramid Matching，SPM)[9]提出将图像均匀分块，然后每个区块里面单独做特征汇聚操作并将所有特征向量拼接起来作为图像最终的特征表达. 空间金字塔匹配的想法非常直观，是金字塔匹配核(Pyramid Matching Kernel，PMK)[27]的图像空间对偶，它操作简单而且性能提

升明显，因而在当前基于词包模型的图像分类框架中成为标准步骤. 实际使用中，在 Caltech 101/256 等数据库上通常使用 1×1、2×2、4×4 的空间分块，因而特征维度是全局汇聚得到的特征向量的 21 倍，在 PASCAL VOC 数据库上，则采用 1×1、2×2、3×1 的分块，因而最终特征表达的维度是全局汇聚的 8 倍.

(4) 使用支持向量机等分类器进行分类. 从图像提取到特征表达之后，一张图像可以使用一个固定维度的向量进行描述，接下来就是学习一个分类器对图像进行分类. 这个时候可以选择的分类器就很多了，常用的分类器有支持向量机、K 近邻、神经网络、随机森林等. 基于最大化边界的支持向量机是使用最为广泛的分类器之一，在图像分类任务上性能很好，特别是使用了核方法的支持向量机. Yang 等人[26]提出了 ScSPM 方法，通过学习过完备的稀疏特征，可以在高维特征空间提高特征的线性可分性，使用线性支持向量机就得到了当时最好的分类结果，大大降低了训练分类器的时间和空间消耗. 随着物体分类研究的发展，使用的视觉单词大小不断增大，得到的图像表达维度也不断增加，达到了几十万的量级. 这样高的数据维度，相比几万量级的数据样本，都与传统的模式分类问题有了很大的不同. 随着处理的数据规模不断增大，基于在线学习的线性分类器成为首选，得到了广泛的关注与应用.

4.2　深度学习模型

深度学习模型[2]是另一类物体识别算法，其基本思想是通过有监督或者无监督的方式学习层次化的特征表达，来对物体进行从底层到高层的描述. 主流的深度学习模型包括自动编码器(Auto-encoder)[28]、受限波尔兹曼机(Restricted Boltzmann Machine，RBM)[29]、深度信念网络(Deep Belief Nets，DBN)[30]、卷积神经网络(Convolutional Neural Netowrks，CNN)[2]、生物启发式模型[31]等.

自动编码器(Auto-encoder)[28]是 20 世纪 80 年代提出的一种特殊的神经网络结构，并且在数据降维、特征提取等方面得到广泛应用. 自动编码器由编码器和解码器组成，编码器将数据输入变换到隐藏层表达，解码器则负责从隐藏层恢复原始输入. 隐藏层单元数目通常少于数据输入维度，起着类似“瓶颈”的作用，保持数据中最重要的信息，从而实现数据降维与特征编码. 自动编码器是基于特征重构的无监督特征学习单元，加入不同的约束，可以得到不同的变化，包括去噪自动编码器(Denoising Autoencoders)[32]、稀疏自动编码器(Sparse Autoencoders)[33]等，这些方法在数字手写识别、图像分类等任务上取得了非常好的结果.

受限玻尔兹曼机[29]是一种无向二分图模型，是一种典型的基于能量的模型(Enery-Based Models，EBM). 之所以称为“受限”，是指在可视层和隐藏层之间有连接，而在可视层内部和隐藏层内部不存在连接. 受限玻尔兹曼机的这种特殊结构，使得它具有很好的条件独立性，即给定隐藏层单元，可视层单元之间是独立的，反之亦然. 这个特性使得它可以实现同时对一层内的单元进行并行 Gibbs 采样. 受限玻尔兹曼机通常采用对比散度(Contrastive Divergence，CD)[30]算法进行模型学习. 受限玻尔兹曼机作为一种无监督的单层特征学习单元，类似于前面提到的特征编码算法，事实上加了稀疏约束的受限玻尔兹曼机可以学到类似稀疏编码那样的 Gabor 滤波器模式.

深度信念网络(DBN)[30]是一种层次化的无向图模型. DBN 的基本单元是 RBM(Restricted Boltzmann Machine)，首先先以原始输入为可视层，训练一个单层的 RBM，然后固定第一层 RBM 权重，以 RBM 隐藏层单元的响应作为新的可视层，训练下一层的 RBM，以此类推. 通过这种贪婪式的无监督训练，可以使整个 DBN 模型得到一个比较好的初始值，然后可以加入标签信息，通过产生式或者判别式方式，对整个网络进行有监督的精调，进一步改善网络性能. DBN 的多层结构，使得它能够学习得到层次化的特征表达，实现自动特征抽象，而无监督预训练过程则极大改善了深度神经网络在数据量不够时严重的局部极值问题. Hinton 等人[30,34]通过这种方式，成功将其应用于手写数字识别、语音识别、基于内容检索等领域.

卷积神经网络(CNN)[2]最早出现在 20 世纪 80 年代，最初应用于数字手写识别，取得了一定的成功. 然而，由于受硬件的约束，卷积神经网络的高强度计算消耗使得它很难应用到实际尺寸的目标识别任务上. Hubel 和 Wiesel[35]在猫视觉系统研究工作的基础上提出了简单、复杂细胞理论，设计出来一种人工神经网络，之后发展成为卷积神经网络. 卷积神经网络主要包括卷积层和汇聚层，卷积层通过使用固定大小的滤波器与整个图像进行卷积，来模拟 Hubel 和 Wiesel 提出的简单细胞. 汇聚层则是一种降采样操作，通过取卷积得到的特征图中局部区块的最大值、平均值来达到降采样的目的，并在这个过

程中获得一定的不变性.汇聚层用来模拟 Hubel 和 Wiesel 理论中的复杂细胞.在每层的响应之后通常还会有几个非线性变换,如 sigmoid、tanh、relu 等,使得整个网络的表达能力得到增强.在网络的最后通常会增加若干全连通层和一个分类器,如 softmax 分类器、RBF 分类器等.卷积神经网络中卷积层的滤波器是各个位置共享的,因而可以大大降低参数的规模,这对防止模型过于复杂是非常有益的,另一方面,卷积操作保持了图像的空间信息,因而特别适合于对图像进行表达.

这里我们将最为流行的词包模型与卷积神经网络模型进行对比,发现两者其实是极为相似的.在词包模型中,对底层特征进行特征编码的过程,实际上近似等价于卷积神经网络中的卷积层,而汇聚层所进行的操作也与词包模型中的汇聚操作一样.不同之处在于,词包模型实际上相当于只包含了一个卷积层和一个汇聚层,且模型采用无监督方式进行特征表达学习,而卷积神经网络则包含了更多层的简单、复杂细胞,可以进行更为复杂的特征变换,并且其学习过程是有监督过程的,滤波器权重可以根据数据与任务不断进行调整,从而学习到更有意义的特征表达.从这个角度来看,卷积神经网络具有更为强大的特征表达能力,因此它在图像识别任务中的出色性能就很容易解释了.

下面我们将以 PASCAL VOC 竞赛和 ImageNet 竞赛为主线,对物体分类的发展进行梳理和分析.

2005 年第一届 PASCAL VOC 竞赛数据库包含了 4 类物体:摩托车、自行车、人、汽车,训练集加验证集一共包含 684 张图像,测试集包含 689 张图像,数据规模相对较小.从方法上来说,词包模型开始在物体分类任务上得到应用,但也存在很多其他的方法,如基于检测的物体分类、自组织网络等.从竞赛结果来看,采用"兴趣点检测-SIFT 底层特征描述-向量量化编码直方图-支持向量机"得到了最好的物体分类性能[36].对数线性模型和 logistic 回归的性能要略差于支持向量机,这也说明了基于最大化边缘准则的支持向量机具有较强的鲁棒性,可以更好地处理物体的尺度、视角、形状等变化.

2006 年玛丽王后学院的 Zhang 等人[37]使用词包模型获得了 PASCAL VOC 物体分类竞赛冠军.与以前不同,在底层特征提取上,他们采用了更多的兴趣点检测算法,包括 Harris-Laplace 角点检测和 Laplacian 块检测.除此以外,他们还使用了基于固定网格的密集特征提取方式,在多个尺度上进行特征提取.底层特征描述除使用尺度不变的 SIFT 特征[16]外,还使用了 SPIN image 特征[38].词包模型是一个无序的全局直方图描述,没有考虑底层特征的空间信息,Zhang 等人采用了 Lazebnik 等人[9]提出的空间金字塔匹配方法,采用 1×1、2×2、3×1 的分块,因而最终特征表达的维度是全局汇聚的 8 倍.另一个与之前不同的地方在于,他们使用了一个两级的支持向量机来进行特征分类,第一级采用卡方核 SVM 对空间金字塔匹配得到的各个词包特征表达进行分类,第二级则采用 RBF 核 SVM 对第一级的结果进行再分类.通过采用两级的 SVM 分类,可以将不同的 SPM 通道结果融合起来,起到一定的通道选择作用.

2007 年来自 INRIA 的 Marszałek 等人[39]获得物体分类冠军,他们所用的方法也是词包模型,基本流程与 2006 年的冠军方法类似.不同在于,他们在底层特征描述上使用了更多的底层特征描述子,包括 SIFT、SIFT-hue、PAS edgel histogram 等,通过多特征方式最大可能保留图像信息,并通过特征编码和 SVM 分类方式发掘有用信息成为物体分类研究者的共识.另一个重要的改进是提出了扩展的多通道高斯核,采用学习线性距离组合的方式确定不同 SPM 通道的权重,并利用遗传算法进行优化.

2008 年阿姆斯特丹大学和萨里大学组成的队伍获得了冠军[40],其基本方法依然是词包模型.有三个比较重要的不同之处,首先是他们提出了彩色描述子来增强模型的光照不变性与判别能力[41];其次是使用软量化编码替代了向量量化编码,由于在实际图像中,图像局部特征常常存在一定的模糊性,即一个局部特征可能和多个视觉单词相似性差别很小,这个时候使用向量量化编码就只使用了距离最近的视觉单词,而忽略了其他同样很相似的视觉单词.为了克服这种模糊性问题,van Gemert 等人提出了软量化编码(又称核视觉词典编码)[7]算法,有效解决了视觉模糊性问题,提高了物体识别的精度.另外,他们还采用谱回归核判别分析得到了比支持向量机更好的分类性能.

2009 年物体分类研究更加成熟,冠军队伍不再专注于多底层特征、多分类器融合,而是采用了密集提取的单 SIFT 特征,并使用线性分类器进行模式分类[42].他们的研究中心放在了特征编码上,采用了混合高斯模型(Gaussian Mixture Model,GMM)和局部坐标编码(Local Coordinate Coding,LCC)[43]两种特征编码方法对底层 SIFT 特征描述子进行

编码，得到了高度非线性的、局部的图像特征表达，通过提高特征的不变性、判别性来改进性能. 另外，物体检测结果的融合，也进一步提升了物体分类的识别性能. 局部坐标编码提出的“局部性”概念，对物体分类中的特征表达具有重要的意义，之后出现的局部线性约束编码(Locality-constrained Linear Coding，LLC)[22]也是基于局部性的思想，得到了“局部的”、“稀疏的”特征表达，在物体分类任务上取得了很好的结果.

2010 年冠军依旧以词包模型为基础，并且融合了物体分割与检测算法[44]. 一方面通过多底层特征、向量量化编码和空间金字塔匹配得到图像的词包模型描述，另一方面，通过使用 Mean shift[45]、过分割[46]、基于图的分割[47]等过分割算法，得到 Patch 级的词包特征表达. 这两种表达作为视觉特征表达，与检测结果以多核学习的方式进行融合. 在分类器方面，除使用了 SVM 核回归外，还提出了基于排他上下文的 Lasso 预测算法. 所谓排他上下文是指一个排他标签集合中至多只能出现一种类别. 排他标签集合的构建使用 Graph Shift 方法，并采用最小重构误差加稀疏约束也即 Lasso 进行预测. 排他上下文作为一种不同于一般共生关系的上下文，高置信度预测可以大大抑制同一排他标签集中其他类别的置信度，改善分类性能.

2011 年冠军延续了 2010 年冠军的基本框架. 来自阿姆斯特丹大学的队伍从最显著窗口对于物体分类任务的作用出发，在词包模型基础上进行了新的探索[48]. 他们发现单独包含物体的图像区域可以得到比整个图像更好的性能，一旦物体位置确定，上下文信息的作用就很小了. 在物体存在较大变化的情况下，部件通常比全局更具有判别性，而在拥挤情况下，成群集合通常要比单个物体更加容易识别. 基于此，他们提出了包含物体部件，整个物体，物体集合的最显著窗口框架. 检测模型训练使用人工标注窗口，预测使用选择性搜索定位. 词包模型和最显著窗口算法融合得到最终的分类结果.

2012 年冠军延续了 2010 年以来的算法框架，在词包模型表达方面，使用了向量量化编码、局部约束线性编码、Fisher 向量编码替代原来的单一向量量化编码[49]. 这里有两个比较重要的改进，一个是广义层次化匹配算法. 考虑到传统的空间金字塔匹配算法在物体对齐的假设下才有意义，而这在实际任务中几乎不能满足，为解决这个问题，他们使用 Side 信息得到物体置信图，采用层次化的方式对局部特征进行汇聚，从而得到更好的特征匹配. 另一个重要的改进是子类挖掘算法，其提出的主要目的是改进类间模糊与类内分散的问题. 基本步骤是：(1)计算样本类内相似度；(2)计算类间模糊性；(3)使用 Graph Shift 算法来检测密集子图；(4)子图向子类的映射.

相比 PASCAL VOC 竞赛，ImageNet 竞赛的图像数据规模更大，类别数更多，对传统的图像分类、检测算法都是一个大的挑战. 下面将近年 ImageNet 竞赛的主流算法也做一个简要介绍.

2010 年冠军由美国 NEC 研究院和 UIUC 获得，其方法基于词包模型，底层特征采用了密集提取的 HOG 和 LBP 特征，特征编码算法使用了局部坐标编码和超向量编码，并且采用了空间金字塔匹配. 最终图像的分类采用了基于平均随机梯度下降的大尺度 SVM. 相比 PASCAL 竞赛算法，这里的算法更多采用了在计算上极为高效的底层特征和编码算法，分类器及其优化也专门针对大规模数据进行了设计，最终获得了 71.8%的 Top 5 分类精度.

2011 年冠军是施乐欧洲研究中心，其基本方法[50]仍旧是基于词包模型，主要改进在 3 个方面：特征编码方法采用 Fisher 向量编码[24]，可以引入更多的高阶统计信息，得到更具判别性的表达；使用乘积量化(Product Quantization，PQ)算法进行特征压缩；分类器使用基于随机梯度下降的线性支持向量机.

2012 年加拿大多伦多大学的 Hinton 教授及其学生 Krizhevsky[34]利用 GPU 在 ImageNet 竞赛上获得了前所未有的成功，他们训练了一个参数规模非常大的卷积神经网络，并通过大量数据生成和 dropout 来抑制模型的过拟合，在大规模图像分类任务上获得了非常好的效果，取得了第一名的成绩，Top 5 分类精度达到了 84.7%，比第二名使用 Fisher 向量编码算法[24]要高大约 10 个百分点，充分显示了深度学习模型的表达能力.

对比 PASCAL 竞赛，ImageNet 竞赛中使用的算法更加简单高效，因而也更加接近实用. 在大规模图像识别场景下，传统图像识别的很多算法和技术面临极大的挑战，包括高计算强度，高内存消耗等，多特征、非线性分类器等这些在 PASCAL 竞赛中广为使用的算法和策略无法在 ImageNet 这样规模的数据库上高效实现. 在性能和效率的权衡中，逐渐被更为简单高效的算法(单特征、特征压缩、线性分类器等)替代. 大数据时代的来临，更激发了数据驱动

的深度学习模型的发展,实现了更高效的特征提取与图像分类,将图像分类的发展推向一个新的高度.

4.3 物体检测

PASCAL VOC 竞赛从 2005 年第一届开始就引入了物体检测任务竞赛,主要任务是给定测试图片预测其中包含的物体类别与外接矩形框.物体检测任务与物体分类任务最重要的不同在于,物体结构信息在物体检测中起着至关重要的作用,而物体分类则更多考虑的是物体或者图像的全局表达.物体检测的输入是包含物体的窗口,而物体分类则是整个图像,就给定窗口而言,物体分类和物体检测在特征提取、特征编码、分类器设计方面很大程度是相通的,如表 3 所示.根据获得窗口位置策略的不同,物体检测方法大致可分为滑动窗口和广义霍夫投票两类方法.滑动窗口方法比较简单,它是通过使用训练好的模板在输入图像的多个尺度上进行滑动扫描,通过确定最大响应位置找到目标物体的外接窗口.广义霍夫投票方法则是通过在参数空间进行累加,根据局部极值获得物体位置的方法,可以用于任意形状的检测和一般物体检测任务.滑动窗口方法由于其简单和有效性,在历年的 PASCAL VOC 竞赛中得到了广泛的使用.特别是 HOG(Histograms of Oriented Gradients)模型、形变部件模型的出现和发展,使得滑动窗口模型成为主流物体检测方法.

表 3 历年 PASCAL VOC 竞赛检测算法

年份	检测策略	底层特征	特征编码	上下文	分类器	融合
2005	滑动窗口、霍夫	SIFT	无	无	线性 SVM	无
2006	滑动窗口	多尺度 HOG	无	无	线性 SVM	多尺度
2007	滑动窗口	多尺度 HOG	无	无	隐 SVM	多尺度,多模型
2008	滑动窗口	多尺度 HOG,密集 SIFT	无	分类结果	线性 SVM、χ^2SVM	多特征,多分类器,分类结果
2009	滑动窗口	多特征	向量量化	无	多级 SVM	多特征,多核学习
2010	滑动窗口	Boosted HOG-LBP	无	全局、空间、类间上下文	隐 SVM、RBF SVM	上下文,分类结果
2011	滑动窗口	Boosted HOG-LBP	无	上下文学习	多分类器	分类结果、上下文学习
2012	滑动窗口	颜色描述子	混合编码	分割	多分类器	分割

与物体分类问题不同,物体检测问题从数学上是研究输入图像 X 与输出物体窗口 Y 之间的关系,这里 Y 的取值不再是一个实数,而是一组“结构化”的数据,指定了物体的外接窗口和类别,是一个典型的结构化学习问题.结构化支持向量机(Structrual SVM,SSVM)[51]基于最大化边缘准则,将普通支持向量机推广到能够处理结构化输出,有效扩展了支持向量机的应用范围,可以处理语法树、图等更一般的数据结构,在自然语言处理、机器学习、模式识别、计算机视觉等领域受到越来越多的关注.隐变量支持向量机(Latent SVM,LSVM)是 Felzenszwalb 等人[52]在 2007 年提出的用于处理物体检测问题,其基本思想是将物体位置作为隐变量放入支持向量机的目标函数中进行优化,以判别式方法得到最优的物体位置.弱标签结构化支持向量机(Weak-Label Structrual SVM,WL-SSVM)是一种更加一般的结构化学习框架,它的提出主要是为了处理标签空间和输出空间不一致的问题,对于多个输出符合一个标签的情况,每个样本标签都被认为是“弱标签”.SSVM 和 LSVM 都可以看做是 WL-SSVM 的特例,WL-SSVM 通过一定的约简可以转化为一般的 SSVM 和 LSVM.条件随机场(Conditional Random Field,CRF)作为经典的结构化学习算法,在物体检测任务上也得到一定的关注.Schnitzspan 等人[53]将形变部件模型与结构化学习结合,提出了一种隐条件随机场模型(latent CRFs),通过将物体部件标签建模为隐藏节点并且采用 EM 算法来进行学习,该算法突破了传统 CRF 需手动给定拓扑结构的缺点,能够自动学习到更为灵活的结构,自动发掘视觉语义上有意义的部件表达.张俊格[54]提出了基于数据驱动的自动结构建模与学习来从训练数据中学习最为合适的拓扑结构.由于一般化的结构学习是一个 NP 难问题,张俊格提出了混合结构学习方案,将结构约束分成一个弱结构项和强结构项.弱结构项由传统的树状结构模型得到,而强结构项则主要依靠条件随机场以数据驱动方式自动学习得到.

下面我们将以历年 PASCAL VOC 物体检测竞赛来探讨物体检测方法的演变与发展.

2005 年物体检测竞赛有 5 支队伍参加,采用的方法呈现多样化[36],Darmstadt 使用了广义霍夫变换,通过兴趣点检测和直方图特征描述方式进行特征表达,并通过广义 Hough 投票来推断物体尺度与位置,该方法在他们参加的几类中都得到了最好的性能.INRIA 的 Dalal 则采用了滑动窗口模型,底层特征使用了基于 SIFT 的描述,分类器使用支持向量机,通过采用在位置和尺度空间进行穷尽搜索,来

确定物体在图像中的尺度和位置，该方法在汽车类别上取得了比广义 Hough 变换更好的性能，但在人、自行车等非刚体类别上性能并不好[36]. 2006 年最佳物体检测算法是 Dalal 和 Triggs 提出的 HOG (Histograms of Oriented Gradients)模型[17]. 他们的工作主要集中于鲁棒图像特征描述研究，提出了物体检测领域中具有重要位置的 HOG 特征. HOG 是梯度方向直方图特征，通过将图像划分成小的 Cell，在每个 Cell 内部进行梯度方向统计得到直方图描述. 与 SIFT 特征相比，HOG 特征不具有尺度不变性，但计算速度要快得多. 整体检测框架依然是滑动窗口策略为基础，并且使用线性分类器进行分类. 这个模型本质上是一个全局刚性模板模型，需要对整个物体进行全局匹配，对物体形变不能很好地匹配处理.

2007 年 Felzenszwalb 等人[52]提出了物体检测领域里程碑式的工作：形变部件模型(Deformable Part-based Model)，并以此取得了 2007 年 PASCAL VOC 物体检测竞赛的冠军. 底层特征采用了 Dalal 和 Triggs 提出的 HOG 特征，但与 Dalal 等人的全局刚体模板模型不同的是，形变部件模型由一个根模型和若干可形变部件组成. 另一个重要的改进是提出了隐支持向量机模型，通过隐变量来建模物体部件的空间配置，并使用判别式方法进行训练优化. 形变部件模型奠定了当今物体检测算法研究的基础，也成为后续 PASCAL VOC 竞赛物体检测任务的基础框架.

2008 年物体检测冠军同样采用了滑动窗口方式[55]. 特征表达利用了 HOG 特征和基于密集提取 SIFT 的词包模型表达. 训练过程对前、后、左、右分别训练独立的模型，并使用线性分类器和卡方核 SVM 进行分类. 测试过程采用了两阶段算法，第一阶段通过滑动窗口方式利用分类器得到大量可能出现物体的位置，第二阶段基于 HOG 和 SIFT 特征对前面一阶段得到的检测进行打分，最后使用非极大抑制算法去除错误检测窗口，并融合分类结果得到最终检测结果. 这里分类信息可以看成是一种上下文信息，这个也是物体检测研究的一个重要内容.

2009 年除了形变部件模型以外，牛津大学视觉几何研究组在滑动窗口框架下，基于多核学习将灰度 PHOW、颜色 PHOW、PHOC、对称 PHOG、SSIM、视觉词典等多种特征进行融合，取得了与形变部件模型相近的效果，获得共同检测冠军[56]. 多核学习是进行多特征、多模型融合的重要策略，可以自动学习多个核矩阵的权重，从而得到最佳的模型融合效果. 考虑到滑动窗口搜索的效率问题，提出了类似级联 Adaboost 方式的多级分类器结构. 第一级分类器采用线性 SVM 分类器以滑动窗口或者跳跃窗口方式快速对图像窗口进行粗分类；第二级采用拟线性 SVM，利用卡方核进行进一步细分类；第三级采用更强的非线性卡方-RBF 分类器，这一步准确度更高但比前面步骤计算代价更大，由于前面两级已经快速滤除大部分备选窗口，这一级可以专注于更难的样本分类.

2010 年中国科学院自动化研究所模式识别国家重点实验室获得了物体检测冠军[57]，其方法是以形变部件模型为基础，对底层 HOG 特征进行了改进，提出了 Boosted HOG-LBP 特征[58]，利用 Gentle Boost 选择出一部分 LBP 特征与 HOG 特征融合，使物体检测结果有了显著提升. 另一个重要改进是采用了多种形状上下文，包括空间上下文、全局上下文、类间上下文. 空间上下文由包含了窗口位置尺度信息的 6 维向量构成，全局上下文包括 20 维的物体分类分数和 20 维的最大窗口分数，其中分类方法采用了 Huang 等人[23]提出的显著性编码、词典关系[59]算法计算词包模型表达. 类间上下文用于建模相邻物体之间的弱空间关系，分别由 20 维的窗口附近最强的 HOG 特征分数和 LBP 特征分数构成. 最终得到 87 维的特征，使用 RBF SVM 进行上下文学习. 该方法在 VOC2010 数据库上取得了 6 项第一，5 项第二，平均精度达到了 36.8%.

2011 年物体检测冠军依然是中国科学院自动化研究所模式识别国家重点实验室[60]，算法上与 2010 年不同之处是针对形变部件模型提出了一种数据分解算法，并引入了空间混合建模和上下文学习[61].

2012 年阿姆斯特丹大学获得物体检测冠军[62]，其方法主要创新在于选择性搜索、混合特征编码、新的颜色描述子、再训练过程. 图像中物体本身构成一种层次结构，通常很难在一个尺度上检测所有物体，因而对图像块进行层次化组织，在每个层次上进行选择搜索，可以有效提升检测的召回率. 考虑到经典的向量量化编码使用小的特征空间分块能够捕获更多图像细节，而丢失了分块内部的细节，而超向量编码和 Fisher 向量量化编码等差异编码方法则可以很好的描述分块内部细节，更大空间分块可以描述

更大范围的图像细节,综合这两种编码模式,提出了混合特征编码算法,将两种编码的优点融合到一起.

5 对物体分类与检测的思考

物体分类与检测的研究在以 PASCAL VOC 竞赛为平台的理论和算法研究上已经取得了一系列的进展,分类模型建立了以词包模型和深度学习模型为基础的体系框架,检测模型则以可形变模型为核心发展出多种方法.在分析目前物体分类和检测算法的基础上,本文接下来对物体分类和检测算法的统一性和差异性进行了讨论,并探讨了物体分类与检测算法发展的方向.

5.1 物体检测和物体分类的统一性

(1)物体检测可以取代物体分类?

物体检测的任务是解决物体所在的位置问题,物体分类的任务是判断物体的种类,从直观上而言,物体检测的隐含信息包括了物体的类别信息,也就是需要事先知道需要定位的物体的类别信息,比如需要检测人,那么就需要先验的给出人的标注信息,以此来判断人的位置,从这个角度而言,物体检测似乎包括了物体分类的步骤,也就是物体检测就能够回答"什么物体在什么地方",但这里有一个误区,其中的"什么物体"是先验给出的,也就是在训练过程中标注出的,并不一定是真实的结果.在模型区分性比较强的情况下,也就是物体检测能给出准确的结果的情况下,物体检测在一定程度上可以回答"什么物体在什么地方",但在真实的世界中,很多情况下模版不能唯一的反映出物体类别的唯一性,只能给出"可能有什么物体在什么地方",此时物体分类的介入就很有必要了.由此可见,物体检测是不能替代物体分类的.

(2)物体检测和物体分类之间的差异性和互补性

以 PASCAL VOC 竞赛为例,从模型的角度而言,物体检测主要采用的是可变的部件模型,更多的关注局部特征,物体分类中主要的模型是词包模型,从两者的处理流程来看,他们利用的信息是不同的,物体检测更多的是利用了物体自身的信息,也就是局部信息,物体分类更多的是利用了图像的信息,也就是全局的信息.他们各有优劣,局部信息考虑了更多的物体结构信息,这使得物体检测和分类的准确性更高,但同时也带来物体分类的鲁棒性不强的问题;全局信息考虑了更多的是图像的全局统计信息,尤其是图像的语义信息,这使得能考虑更多的信息来进行判断,但信息量的增加可能带来准确度的提高,也可能由于冗余降低分类的性能,但是从统计意义而言,其鲁棒性是能够得到一定的提高的.由此可见,物体检测和物体分类之间存在着较大的差异性,同时也就说明存在着比较大的互补性.

5.2 物体分类与检测的发展方向

物体分类任务要确定图像中是否包含物体,全局表达更关键;物体检测任务则要确定图像中物体的位置和尺度,物体结构更为关键.因此,物体分类检测的研究也主要有两种思路:

(1)专注于学习结构,即结构化学习.观察变量与其他变量构成结构化的图模型,通过学习得到各个变量之间的关系,结构包括有向图模型(贝叶斯网络)、无向图模型(马尔科夫网络).结构化学习通常变量具有显式的物理意义,变量之间的连接也具有较强的因果关系,解释性较好.

(2)专注于学习层次化表达,即深度学习.深度学习从人脑的层次化视觉处理和函数表达理论出发,采用层次化特征表达的思想来进行特征从底层到高层语义的提取.深度学习专注于表达的学习,也即更注重一个输入得到的相应输出,对中间的特征变换缺少自然的解释,更像一个黑盒系统.

两条思路各有侧重,但并不是互相独立的.在这两条发展线路的基础上,建立更为统一的物体识别框架,同时处理物体分类与检测任务,是一个更加值得研究的方向.如何利用物体检测和物体分类之间的互补性去构建统一的物体识别框架是计算机视觉和视觉认知领域的研究热点,也是视觉认知计算模型研究的重点之一[63].

5.3 结构化学习存在的难点与挑战

(1)模型表达问题.对于一个特定问题,选择什么样的模型,如有向图模型、无向图模型,模型如何进行参数化,都是值得研究的.

(2)模型学习问题.在给定模型表达后,如何从给定数据中学习模型的参数,是结构化学习中的一个核心问题.目前通常有基于概率的学习方法,如最大似然估计、最大后验估计等,也有基于最小化损失函数的方法.不同的方法,在学习的效率,准确性上都具有差异,研究快速有效的学习算法,具有特别重要的价值.

(3)模型推断问题.给定学习好的模型,进行快速、准确的模型推断是至关重要的.目前经典的方法包括消息传播算法、变分推断算法、采样算法等.不同方法在速度、准确度上各有差异.研究大规模图模型,实现人类视觉系统快速识别人脸[64]那样的快速准确推断,是一个重要研究方向.

5.4　层次化学习(深度学习)存在的难点与挑战

在大数据时代，海量的图像、视频数据绝大多数是没有标签的，大量进行标注也是不现实的. 从大量的没有标签的图像数据中自动挖掘知识，无疑有着重要的意义. Google Brain 计划[65-66]也验证了数据驱动的自主学习的可行性与有效性. 但目前深度学习还存在一些难点和挑战.

(1) 解释性差. 层次化表达在视觉皮层理论和函数论等方面具有其理论依据[67-68]，然而，在实际应用中，学习到的模型通常没有很好的解释性. 第一层网络可以通过可视化的方式进行直接查看，在大多数视觉数据中，第一层学习到的是类似 Gabor 的滤波器，可以实现基本的边缘检测. 然而，对于更高层的特征，通常很难直观的查看其学习到的是什么. 研究有效的高层特征解释方式，无疑对于深度学习的发展具有非常重要的意义.

(2) 模型复杂度高，优化困难. 神经网络的容量没有上限，表达能力非常强，这是它的一个重要的优点. 另一方面也对模型的优化造成了非常大的困难. 网络越复杂，模型的能量面越高低不平，到处是极小点. 研究模型初始化方式、优化算法，提高神经网络的判别能力，是深度学习的一个重要研究内容.

(3) 计算强度高. 目前虽然每层是高度并行化的前馈网络，但是计算强度还是比较高，需要采用 GPU 等硬件来完成. 对于一个刺激信号，人脑中绝大多数细胞是处于不活动状态，只有相关的细胞才会有活动，这是一种非常经济的响应形式. 而对于深度学习，输入一个视觉信号，所有的神经元都会进行计算，人为加的一些稀疏约束只是会使某些神经元输出为 0，但不代表该神经元“处于不活动”状态. 这方面是将来建立庞大学习网络时实现实时推理的一个可行思路.

(4) 模型缺少结构约束. 深度学习模型通常只对网络的“输入-输出”进行建模，却缺少必要的结构先验的约束. 例如，对人脸关键点可以采用卷积神经网络进行回归，网络学习到的是一种隐式的“输入-输出”结构，却完全没有加入显式的结构先验，包括预测输出的位置点处的表观特征. 这个 问题的直接后果就是单个网络尽管可以做到任意的复杂度，却无法得到很高的精度，很多检测错误看起来是非常简单的：本来应该落在具有明显特征的嘴角处，却落在了嘴角旁边的脸部区域. 为了克服这个问题，就需要采用从粗到细，从全局到局部的策略，级联多个网络来不断纠正网络预测.

在大数据时代，海量视频数据所带来的纷繁复杂的易变性(variability)将给传统的特征学习方法带来巨大挑战. 而深度学习模型天然的强大数据表达能力，无疑将会对大数据背景下的整个视觉的研究产生极大的影响，也必然会将图像物体检测、分类的研究推向新的高度. 当然，目前深度学习模型还存在着解释性差、模型复杂度高，优化困难、计算强度高等诸多问题，这些都需要研究者们进一步的思考. 例如，将显式结构先验嵌入深度学习模型中，可以有效降低网络参数空间的规模，减少局部极值的问题，从而可以更加有效地解决检测、分割等任务.

6　结　论

物体分类与检测在计算机视觉研究中具有重要的理论意义和实际应用价值，同时目前也存在诸多困难与挑战. 本文以计算机视觉物体识别算法竞赛 PASCAL VOC 为主线，对物体分类与检测历年最佳算法的发展进行了详尽的阐述，强调了表达学习和结构学习分别在物体分类和物体检测中的重要意义. 以此为基础，本文还讨论了物体分类与检测的统一性与差异性，对物体分类与检测的发展方向进一步思考，从基于深度学习的表达学习和结构学习两个方向进行了分析与展望，我们有理由相信，物体分类和物体检测算法的统一发展必然会促进图像物体识别的进展.

致　谢　本文得到国家自然科学基金委员会、国家科学技术部等机构的支持，中国科学院自动化研究所 PASCAL VOC 竞赛团队的成员也付出了不懈努力，在此一并感谢！最后，谨以此文纪念英国利兹大学的 Everingham 博士，他和 Zisserman 等人在推动 PASCAL VOC 视觉物体识别竞赛方面做出了重要的贡献！

参　考　文　献

[1] Marr D. Vision: A Computational Investigation Into the Human Representation and Processing of Visual Information. Cambridge: The MIT Press, 2010

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

[3] Ferrari V, Jurie F, Schmid C. From images to shape models for object detection. International Journal of Computer Vision, 2009, 87(3): 284-303

[4] Latecki L J, Lakamper R, Eckhardt U. Shape descriptors for non-rigid shapes with a single closed contour//Proceedings of

the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Hilton Head, USA, 2000, 1: 424-429

[5] Krizhevsky A. Learning Multiple Layers of Features from Tiny Images[M. S. dissertation]. University of Toronto, 2009

[6] Torralba A, Fergus R, Freeman W T. 80 million tiny images: A large dataset for non-parametric object and scene recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2008, 30(11): 1958-1970

[7] Li Fei-Fei, Fergus R, Perona P. Learning generative visual models from few training examples: An incremental Bayesian approach tested on 101 object categories//Proceedings of the Computer Vision and Pattern Recognition(CVPR), Workshop on Generative-Model Based Vision. Washington, USA, 2004: 178

[8] Griffin G, Holub A D, Perona P. The Caltech 256. Caltech Technical Report CNS-TR-2007-001

[9] Lazebnik S, Schmid C, Ponce J. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). New York, USA, 2006: 2169-2178

[10] Li Fei-Fei, Perona P. A Bayesian hierarchical model for learning natural scene categories//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Washington, USA, 2005: 524-531

[11] Everingham M, Van Gool L, Williams C K I, et al. Introduction to PASCAL VOC 2007//Proceedings of the Workshop on PASCAL Visual Object Classes Challenge. Rio de Janeiro, Brazil, 2007

[12] Deng Jia, Dong Wei, Richard S, et al. ImageNet: A large-scale hierarchical image database//Proceedings of the IEEE Computer Vision and Pattern Recognition (CVPR). Miami, USA, 2009: 248- 255

[13] Xiao Jian-Xiong, Hays J, Ehinger K, et al. SUN database: Large-scale scene recognition from abbey to zoo//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). San Francisco, USA, 2010: 3485-3492

[14] Torralba A, Efros A A. Unbiased look at dataset bias//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Colorado Springs, USA, 2011: 1521-1528

[15] Csurka G, Dance C R, Fan Li-Xin, et al. Visual categorization with bags of keypoints//Proceedings of the Workshop on Statistical Learning in Computer Vision, Proceedings of the 8th European Conference on Computer Vision. Prague, Czech, 2004: 1-22

[16] Lowe G D. Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision, 2004, 60: 91-110

[17] Dalal N, Triggs B. Histograms of oriented gradients for human detection//Proceedings of the Computer Vision and Pattern Recognition (CVPR). San Diego, USA, 2005: 886-893

[18] Ojala T, Pietikäinen M, Mäenpää T. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2002, 24(7): 971-987

[19] Sivic J, Zisserman A. Video Google: A text retrieval approach to object matching in videos//Proceedings of the IEEE International Conference on Computer Vision. Madison, USA, 2003: 1470-1477

[20] van Gemert J C, Geusebroek J-M, Veenman C J, Smeulders A W M. Kernel codebooks for scene categorization//Proceedings of the European Conference on Computer Vision (ECCV). Marseille, France, 2008: 696-709

[21] Olshausen B A, Fieldt D J. Sparse coding with an overcomplete basis set: A strategy employed by v1. Vision Research, 1997, 37: 3311-3325

[22] Wang Jinjun, Yang Jianchao, Yu Kai, et al. Locality-constrained linear coding for image classification//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). San Francisco, USA, 2010: 3360-3367

[23] Huang Yongzhen, Huang Kaiqi, Yu Yinan, Tan Tieniu. Salient coding for image classification//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Colorado, USA, 2011: 1753-1760

[24] Perronnin F, Sánchez J, Mensink T. Improving the fisher kernel for large-scale image classification//Proceedings of the European Conference on Computer Vision (ECCV). Crete, Greece, 2010, 6314: 143-156

[25] Zhou Xi, Yu Kai, Zhang Tong, Huang T S. Image classification using super-vector coding of local image descriptors//Proceedings of the European Conference on Computer Vision (ECCV). Berlin, Germany, 2010: 141-154

[26] Yang Jianchao, Yu Kai, Gong Yihong, Huang T. Linear spatial pyramid matching using sparse coding for image classification//Proceedings of the IEEE International Conference on Computer Vision (ICCV). Miami, USA, 2009: 1794-1801

[27] Grauman K, Darrell T. The pyramid match kernel: Discriminative classification with sets of image features//Proceedings of the IEEE International Conference on Computer Vision (ICCV). Beijing, China, 2005: 1458-1465

[28] Bourlard H, Kamp Y. Auto-association by multilayer perceptrons and singular value decomposition. Biological, Cybernetics, 1988, 59: 291-294

[29] Smolensky P. Chapter 6: Information processing in dynamical systems: Foundations of harmony theory//Processing of the Parallel Distributed: Explorations in the Microstructure of Cognition, Volume 1: Foundations. MIT Press, 1986

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

[31] Huang Yongzhen, Huang Kaiqi, Tao Dacheng, et al. Enhanced biologically inspired model for object recognition. IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics, 2011, 41(6): 1668-1680

[32] Vincent P, Larochelle H, Bengio Y, Manzagol P A. Extracting and composing robust features with denoising autoencoders//Proceedings of the 25th International Conference on Machine Learning (ICML). Helsinki, Finland, 2008: 1096-1103

[33] Coates A, Ng A Y, Lee H. An analysis of single-layer networks in unsupervised feature learning. Journal of Machine Learning Research, 2011, 15(1): 215-223

[34] Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks//Proceedings of the Advances in Neural Information Processing Systems (NPIS). Lake Tahoe, USA, 2012: 1106-1114

[35] Hubel D H, Wiesel T N. Receptive fields of single neurons in the cat's striate cortex. Journal of Physiology, 1959, 148: 574-591

[36] Everingham M, Zisserman A, Williams C, et al. The 2005 PASCAL visual object classes challenge//Proceedings of Workshop on the First PASCAL Challenges. Graz, Austria, 2006

[37] Zhang Jianguo, Marszałek M, Lazebnik S, Schmid C. Local features and kernels for classification of texture and object categories: A comprehensive study. International Journal of Computer Vision, 2007, 73(2): 213-238

[38] Svetlana L, Schmid C, Ponce J. A sparse texture representation using local affine regions. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2005, 27(8): 1265-1278

[39] Marszałek M, Schmid C, et al. Learning representations for visual object class recognition//Proceedings of the Workshop on PASCAL VOC. Rio de Janeiro, Brazil, 2007

[40] van de Sande K, Uijlings J, Li Xi-Rong, et al. Surrey UVA_SRKDA method//Proceedings of the Workshop on PASCAL VOC. Marseille, France, 2008: 37-46

[41] van de Sande K E A, Geversand T, Snoek C G M. Evaluation of color descriptors for object and scene recognition//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Anchorage, USA, 2008: 1-8

[42] Gong Yihong, Huang Thomas, Lv Fengjun, et al. Image classification using Gaussian mixture and local coordinate coding//Proceedings of the Workshop on Pascal VOC. Kyoto, Japan, 2009: 1-8

[43] Yu Kai, Zhang Tong, Gong Yihong. Nonlinear learning using local coordinate coding//Proceedings of the Advances in Neural Information Processing Systems (NPIS). Vancouver, Canada, 2009

[44] Yan Shui-Cheng, Huang Zhong-Yang, Chen Qiang, et al. Boosting classification with exclusive context//Proceedings of the Workshop on Visual Recognition Challenge. Crete, Greece, 2010: 1-8

[45] Comanicu D, Meer P. Mean shift: A robust approach toward feature space analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2002, 24(2): 603-619

[46] Achanta R, Shaji A, Smith K, et al. SLIC super-pixels compared to state-of-the-art super-pixel methods. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2012, 34(11): 2274-2282

[47] Felzenszwalb P F, Huttenlocher D P. Efficient graph-based image segmentation. International Journal of Computer Vision, 2004, 25(2): 167-181

[48] Uijlings J, van de Sande K, Smeulders A, et al. The most telling window for image classification//Proceedings of the Workshop on PASCAL Visual Object Classes Challenge. Barcelona, Spain, 2011: 1-8

[49] Chen Qiang, Song Zheng, Hua Yang, et al. Hierarchical matching with side information for image classification//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Providence, USA, 2012: 3426-3433

[50] Perronnin F, Sánchez J. Compressed fisher vectors for LSVRC//Proceedings of the Workshop on PASCAL VOC/ImageNet, ICCV. Barcelona, Spain, 2011: 15-23

[51] Tsochantaridis I, Joachims T, Hofmann T, Altun Y. Large margin methods for structured and interdependent output variables. Journal of Machine Learning Research, 2006, 6(2): 1453-1484

[52] Felzenszwalb P, McAllester D, Ramanan D. A discriminatively trained, multiscale, deformable part model//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Anchorage, Alaska, USA, 2008: 1-8

[53] Schnitzspan P, Roth S, Schiele B. Automatic discovery of meaningful object parts with latent CRFs//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). San Francisco, USA, 2010: 121-128

[54] Zhang Jun-Ge. Object Detection Based on Visual Structure [Ph. D. dissertation]. Institute of Automation, Chinese Academy of Sciences, Beijing, 2013(in Chinese)
(张俊格. 基于视觉结构表达与建模的物体检测研究[博士学位论文]. 中国科学院自动化研究所, 北京, 2013)

[55] Harzallah H, Schmid C, Jurie F, Gaidon A. Classification aided two stage localization//Proceedings of the Workshop on PASCAL Visual Object Classes Challenge. Marseille, France, 2008: 15-23

[56] Vedaldi A, Gulshan V, Varma M, Zisserman A. Multiple kernels for object detection//Proceedings of the IEEE International Conference on Computer Vision (ICCV). Kyoto, Japan, 2009: 606-613

[57] Yu Yinan, Zhang Junge, Huang Yongzhen, et al. Object detection by context and boosted HOG-LBP//Proceedings of the Workshop on PASCAL VOC. Crete, Greece, 2010: 24-32

[58] Zhang Junge, Huang Kaiqi, Yu Yinan, Tan Tieniu. Boosted local structured HOG-LBP for object localization//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Colorado, USA, 2011: 1393-1400

[59] Huang Yongzhen, Huang Kaiqi, Wang Chong, Tan Tieniu. Exploring relations of visual codes for image classification//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Colorado, USA, 2011: 1649-1656

[60] Zhang Junge, Yu Yinan, Huang Yongzhen, et al. Object detection based on data decomposition, spatial mixture modeling and context//Proceedings of the Workshop on PASCAL VOC. Barcelona, Spain, 2011: 24-32

[61] Zhang Junge, Huang Yongzhen, Huang Kaiqi, et al. Data decomposition and spatial mixture modeling for part based model//Proceedings of the Asian Conference on Computer Vision (ACCV). Daejeon, Korea, 2012: 123-137

[62] van de Sande K, Uijlings J, Snoek C, Smeulders A. Hybrid coding for selective search//Proceedings of the Workshop on PASCAL VOC. Florence, Italy, 2012: 1-8

[63] Huang Kai-Qi, Tan Tie-Niu. Review on computational model for vision. Pattern Recognition and Artificial Intelligence, 2013, 26(10): 951-958(in Chinese)
(黄凯奇, 谭铁牛. 视觉认知计算模型综述. 模式识别与人工智能, 2013, 26(10): 951-958)

[64] Linkenkaer-Hansen K, Palva J M, Sams M, et al. Face-selective processing in human extrastriate cortex around 120 ms after stimulus onset revealed by magneto- and electroencephalography. Neuroscience Letters, 1998, 253(3): 147-150

[65] Dean J, Corrado G, Monga R, et al. Large scale distributed deep networks//Proceedings of the Advances in Neural Information Processing Systems (NPIS). Lake Tahoe, USA, 2012: 1232-1240

[66] Le Q V, Ranzato M A, Monga R, et al. Building high-level features using large scale unsupervised learning//Proceedings of the International Conference on Machine Learning (ICML). Edinburgh, UK, 2012: 107-114

[67] Bengio S, Weston J, Grangier D. Label embedding trees for large multi-class tasks//Proceedings of the Advances in Neural Information Processing Systems (NPIS). Vancouver, Canada, 2010: 163-171

[68] Bengio Y. Learning Deep Architectures for AI. Foundations and Trends(r) in Machine Learning. Now Publisher Inc, 2009

HUANG Kai-Qi, born in 1977, Ph. D., professor. His research interests include computer vision, pattern recognition and visual surveillance.

REN Wei-Qiang, born in 1985, Ph. D. candidate. His research interests include computer vision and pattern recognition.

TAN Tie-Niu, born in 1964, Ph. D., professor, member of Chinese Academy of Sciences. His research interests include biometrics, visual surveillance, Internet information forensics and security.

Background

Object classification and detection are two of the most essential problems in computer vision. They are the basis of many other complex vision problems, such as image segmentation, visual object tracking, scene understanding, and action analysis. Due to the large variants of view, scale, illumination, deformation of objects, object classification and detection are still challenging tasks in real environments.

In this paper, we try to give a review of image object classification and detection based on PASCAL VOC challenge, which is generally acknowledged as a public evaluation for object recognition. From 2005 to 2012, PASCAL VOC challenge is held annually with three main competition tasks: object classification, object detection and segmentation. The methods adopted in VOC challenges can be viewed as the state-of-the-art object recognition algorithm, which are introduced in this paper. We summarize the difficulties and challenges in the development of object recognition from three levels: the instance level, the category level and the semantic level. Then we review the yearly achievements in the study of object classification and detection. For object classification, bag-of-words (BoW) and deep learning models are two of the most successful models. For object detection, deformable part based model (DPBM) is the state-of-the-art method, with many extensions, such as grammar model, context, etc. Most of the methods used in competitions are combinations of many different algorithms, with lots of carefully chosen hyper-parameters.

Based on the analysis of recent progresses in PASCAL VOC competition, we discuss the future development directions of object classification and detection, from the view of feature representation learning and model structure learning.

Our group wons the PASCAL VOC 2010 and 2011 object detection competition and got 2nd place in PASCAL VOC 2010 and 2011 object classification competition. The methods used for object detection is deformable part based model with context and boosted HOG-LBP features. Salient coding, dictionary learning and code relations are used in object classification implementation.

This work is funded by the National Basic Research Program of China (Grant No. 2012CB316302), National Natural Science Foundation of China (Grant No. 61322209), the National Key Technology R&D Program (Grant No. 2012BAH07B01).