浙江大学

计算机视觉课程报告
（2018-2019 春）

题目　基于计算机视觉的图像物体分类算法综述

姓名　周金华

学号　21821084

学科、专业　软件工程

任课教师　王东辉

基于计算机视觉的图像物体分类算法综述

1. 引言

物体分类是计算机视觉、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，物体分类已经应用于人们日常生活的方方面面，计算机自动分类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。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；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。物体分类的研究，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）。鉴于物体分类在计算机视觉领域的重要地位，研究鲁棒、准确的物体分类算法，无疑有着重要的理论意义和实际意义。

本文从物体分类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研究中存在的困难与挑战进行了阐述。然后对物体检测和分类方面的主流算法进行梳理总结：基于表达学习和结构学习。在此基础上，本文对物体分类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. 物体分类的难点

物体分类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。

（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，比如一头驴和一头骡子，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。即同样的图像，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 PACALVOC 竞赛[2]为主线对物体分类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、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 PASCALVOC 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。

3.1 基于词包模型的分类

词包模型是物体分类算法的基本框架，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。

词包模（Bagofwords）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。Csyrka 等人[3]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。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。

（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。最简单的特征编码是向量化编码。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。

(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。

(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。

3。2 深度学习模型

深度学习模型[4]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、受限波尔兹曼机、深度信念网络、卷积神经网络、生物启发式模型等。

自动编码器是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。

受限玻尔兹曼机是一种无向二分图模型，是一种典型的基于能量的模型。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。

深度信念网络是一种层次化的无向图模型。ＤＢＮ的基本单元是ＲＢＭ，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。

卷积神经网络最早出现在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。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 Huble 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 Huble 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，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 sofmax 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。

4. 对物体分类的思考

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；因此，物体分类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。

参考文献

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

[2] Everingham M, Van Gool L, Williams C K I, et al. Introduction to PASCAL VOC, 2007

[3] Csurka G, Dance CR, Fan Li-Xin ,et,al. Visual categorization with bags of keypoints. 2004: 1-22

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