计算机视觉模式识别综述

周韧哲 (181220076、zhourz@smail.nju.edu.cn)

摘　要: 计算机视觉模式识别是人工智能领域的一个重要的方向，历经几十年的发展已经成为了一个成熟的学科，本文首先介绍了计算机视觉模式识别的定义，然后深入介绍了视觉模式识别的主要任务和技术进展，由此思考了当下计算机视觉模式识别所面临的问题和挑战，并探讨了计算机视觉模式识别与人类智能的异同。

关键词: 计算机视觉，模式识别，人工智能

1　引言

计算机视觉（Computer vision）是一门研究如何使机器“看”的科学，更进一步的说，就是指用摄影机和计算机代替人眼对目标进行识别、跟踪和测量等机器视觉，并进一步做图像处理，用计算机处理成为更适合人眼观察或传送给仪器检测的图像[1] 。计算机视觉赋予机器类人甚至超人的视觉感知和认知能力，是人工智能的基础问题。计算机视觉任务包括：获取，处理，分析和理解数字图像，以及从现实世界中提取高维数据，以便产生数字或符号信息。计算机视觉涉及到一个基本概念：模式识别。模式识别（Pattern Recognition）旨在将数据（模式）分类为多个类别，是一种自动判别和分类的理论[2] 。什么是模式呢？援引 1985 年 Satoshi Watanabe 对模式的定义：模式是混沌的对立面，一个可以被命名的实体（entity）。可以理解为，模式可以是任何需要识别的实体。在计算机科学领域，更精确的定义是：为了让机器执行和完成识别任务，必须对分类识别对象进行科学的抽象，建立其数学模型，用以描述和代替识别对象，这种对象的描述就是模式。模式的示例包括：图像中的像素，2D 或 3D 形状，打字或手写字符，个体的步态、手势、指纹、足迹、人脸，心电图时间序列，建筑物，动物的形状等等。在计算机视觉方面的模式识别，主要任务有人脸识别、目标检测、图像分割、目标分类等。

2　进展

“模式识别”被广泛使用和形成一个特定的领域是在 20 世纪 60 年代后，1966 年，IBM 在波多黎各组织召开了第一次以“模式识别”为题的学术会议。20 世纪 70 年代，与模式识别有关的教材出版，1972 年，第一届国际模式识别大会（ICPR）召开，在 1978 年召开的第四届大会上成立了国际模式识别协会（IAPR）。80 年代后，模式识别领域在统计学习理论、集成学习、贝叶斯学习和神经网络等方向上发展迅速。

早期的计算机视觉领域的研究工作包括了 60 年代 MIT 的 Roberts 通过计算机程序从数字图像中提取出诸如立方体、楔形体、棱柱体等多面体的三维结构，70 年代 MIT 的 AI 实验室提出了计算机视觉领域的一个重要的理论框架：计算视觉理论。后来计算机视觉与机器学习方法结合。在 21 世纪初期，随着基于统计学习的模式识别方法的快速发展，基于学习的视觉成为了计算机视觉的主流研究方向，尤其是以深度学习加速的计算机视觉在多个领域和任务上都获得了比传统方法好得多的结果。

2009 年，为了建设更好的训练数据集，当时在普林斯顿大学任教的李飞飞和其团队建立了一个超大规模

的项目：ImageNet。ImageNet 由 320 万个标记图像组成，分为 5247 个类别。同年，李飞飞和同事一起，根据数据集撰写了 5 篇论文，探讨了算法如何解释这样大量的数据。第一篇论文后来成为如何处理几千种图像的基准。2010 年起举办的“ImageNet Competition for Object Recognition”竞赛，使得 ImageNet 数据集很快成为图像分类算法在处理当时最复杂视觉数据集时的基准。该竞赛加速推动了计算机视觉模式识别的进展。

以下基于计算机视觉模式识别中的一些基本任务来介绍领域的进展。

2.1 人脸识别

人脸识别属于图像分类任务，作为计算机视觉模式识别的一个极其重要的子领域，向来就是学者的热门研究方向：让计算机理解“人脸”可以说是计算机视觉的一个无可避免的课题。在过去的 20 年里，研究者提出很多人脸识别的方法。大多数研究者运用了整张人脸来做识别，基本技术路线都是提取人脸特征然后进行比对。

一种方法是基于关键点的高维人脸特征提取方法[3] 。高维特征提取的关键在于对人脸的关键点的定位，然后对倾斜的人脸进行矫正，那么标定人脸的关键点的位置（眼睛，鼻子，嘴巴等关键点）就是关键，高维的含义是提取了多个关键点和多个尺度的特征。

还有一种方法是特征脸（Eigenface）的技术。特征脸的方法是 90 年代初期由 Turk 和 Pentland 提出的目前最流行的算法之一，具有简单有效的特点，也称为基于主成分分析（PCA）的人脸识别方法[4] 。特征脸技术的基本思想是：从统计的观点，寻找人脸图像分布的基本元素，即人脸图像样本集协方差矩阵的特征向量，以此近似地表征人脸图像。这些特征向量称为特征脸。实际上，特征脸反映了隐含在人脸样本集合内部的信息和人脸的结构关系。自 1991 年特征脸技术诞生以来，研究者对其进行了各种各样的实验和理论分析，目前，改进的特征脸算法是主流的人脸识别技术，也是具有最好性能的识别方法之一。

与上述完全人脸识别不同的是，部分人脸的信息缺失会导致一些特征无法提取。Renliang Weng 等人利用关键点位置信息和纹理信息来对人脸进行识别[5] ，关键在于对人脸的校正。其方法是对图像不断地进行仿射变换，使完全脸和部分脸的关键点之间距离和以及纹理特征之间的距离和达到全局最小，从而完成对图像的校正。

在计算机的算力得到极大提升的深度学习时代，人脸识别准确率达到了极点，2012 年 AlexNet 在 ImageNet 数据集上大放异彩，使得卷积神经网络（CNN）得到了研究者的广泛关注。 LFW 数据集（Labeled Faces in the Wild）是目前用得最多的人脸图像数据库。该数据库共 13233 幅图像，其中 5749 个人，其中 1680 人有两幅及以上的图像，4069 人只有一幅图像，采集的是自然条件下人脸图片，目的是提高自然条件下人脸识别的精度。国内的旷视科技（face++）从网络上搜集了 5 千万张人脸图片用于训练深度卷积神经网络模型，开发了多分类的十层深度神经网络，最后一层在训练阶段设置为监督学习，而之前的隐藏层输出作为输入图像的特征，使用了上文提到的主成分分析方法进行了人脸的特征缩减，通过一个简单的 L2 范数（衡量向量或矩阵的距离函数）来测量两个图像之间的相似度。该模型在 LFW 数据集上准确率非常高，达到了 0.995 的准确率[6] 。

传统的人脸识别流程是：检测、对齐、表示、分类。Deepface 使用了 3D 的人脸建模来重现对齐和表示这两步，最终从一个 9 层的深度神经网络中得到了人脸的表达。网络参数超过了 120000000 个，在有 4000 多个不同的人，总计 440 万张带标记的人脸数据库上训练。这种在大型数据库中基于模型进行准确的对齐并用神经网络训练学习到的人脸表达，可以很好地推广到非受限环境下的人脸表达。Deepface 模型在 LFW 上达到了 97.35%的人脸验证精度，逼近了人类的水平[7] 。

2.2 物体检测

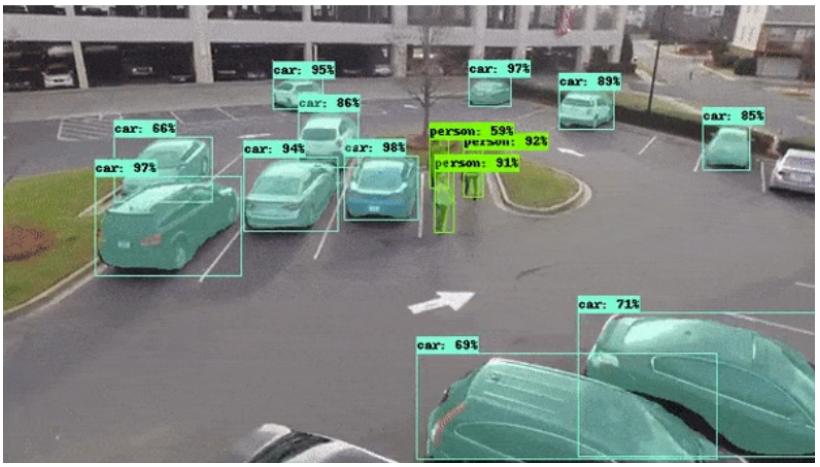
物体检测任务可以描述为：给定一副图像，找出图像中的物体，并定位出这些物体的位置。一个标准的物体检测算法需要用矩形框将图片中的物体位置标出，并给出物体的具体类别。在计算机视觉模式识别中，物体检测是最基本的视觉问题之一，也是物体跟踪、行为分析等其它高层视觉任务的基础。物体检测在现实场景中也有着非常重要的应用，如视频监控、无人驾驶等。在实际应用中，物体检测往往面临诸多挑战，如

图像的光照条件、拍摄视角及距离、物体自身的非刚性形变以及遮挡等因素都会给检测算法带来极大困难。为解决上述问题，学者在特征提取、模型设计及分类器学习等方面做了大量研究工作。

传统经典物体检测算法使用滑动窗口的方式，将不同尺度大小的窗口滑动到待检测图像的不同位置，然后提取图像特征并判断窗口内是否含有待检测物体。受限于特征的表达能力及分类器的判别能力，传统检测算法在现实场景的鲁棒性能并不理想。自 2012 年深度学习方法在 ImageNet 竞赛上斩获成功后，基于深度学习的物体检测算法开始大放光彩。深度神经网络能有效地从原始海量数据集里学习层级化的特征表达，且不需要太多人工干预。

基于区域候选模型先通过区域候选产生感兴趣区域，再对感兴趣区域进行特征提取、判别及检测等操作。Girshick 等人在 2014 提出的 R-CNN 是基于区域候选模型算法的开山之作，利用选择搜索产生候选区域，避免了低效的滑动窗口操作，接着利用卷积神经网络提取特征，最后通过分类器进行判别并使用回归模型回归检测框的位置[8] 。

基于区域候选模型通常采取两阶段检测的策略，即先提取候选区域，再进一步进行检测操作。而基于回归模型通常采取单阶段检测的策略，使用回归的思想，直接对给定图片的多个位置上回归目标的边框及类别，因而有着更快的检测速度。Redmon 等人在 Faster R-CNN 提出的 YOLO 框架使用卷积神经网络直接回归目标物体的置信度、类别及边框坐标信息，将目标检测任务转换成回归问题，大大提升了检测速度，使得其算法能够应用在实时场景中[9] 。



图表 1：自动驾驶中的物体检测

2.3 图像分割

图像分割是计算机视觉研究中的一个经典难题，所谓图像分割是指根据灰度、彩色、空间纹理、几何形状等特征把图像划分成若干个互不相交的区域，使得这些特征在同一区域内表现出一致性或相似性，而在不同区域间表现出明显的不同。简单的说就是在一副图像中，把目标从背景中分离出来。

基于阈值的分割方法的基本思想是基于图像的灰度特征来计算一个或多个灰度阈值，并将图像中每个像素的灰度值与阈值作比较，最后将像素根据比较结果分到合适的类别中。阈值法特别适用于目标和背景占据不同灰度级范围的图像，首先将图像划分成背景区域与目标区域， 根据图像灰度直方图信息获取分割阈值实现分割。基于阈值的分割方法易受噪声影响，很难找到合适的分割阈值[10] 。

基于区域的分割方法是以直接寻找区域为基础的分割技术，利用局部空间信息进行区域分割，将具有相似特征的像素组成一个区域。基于区域提取方法有两种基本形式：一种是区域生长，从单个像素出发，逐步合并以形成所需要的分割区域；另一种是从全局出发，逐步切割至所需的分割区域[10] 。

基于神经网络的分割方法的基本思想是通过训练多层感知机来得到线性决策函数，然后用决策函数对像素进行分类来达到分割的目的。这种方法需要大量的训练数据。神经网络存在巨量的连接，容易引入空间信息，能较好地解决图像中的噪声和不均匀问题。选择何种网络结构是这种方法要解决的主要问题[10] 。

3 挑战

自从 21 世纪至今，计算机视觉模式识别突飞猛进，一路高歌，模型准确率甚至已经超过了人类。例如对于图像分类，在上文所提到的 ImageNet 数据集上，目前先进算法的表现就已经超过了人类。视觉模式识别也已经应用在了如视频监控、自动驾驶和智能医疗等方面，诞生了诸如旷视、商汤、依图、云从科技等 AI 独角兽公司。视觉模式识别巨大进展的背后推动力是深度学习。深度学习的成功主要得益于三个方面：大规模数据集的产生、强有力的模型的发展以及可用的大量计算资源。对于各种各样的视觉模式识别任务，精心设计的深度神经网络已经远远超越了以前那些基于人工设计的图像特征的方法。尽管到目前为止深度学习在计算机视觉模式识别方面已经取得了巨大成功，但在它进一步广泛应用之前，仍然有很多挑战需要我们去面对。

一个重要的挑战是，怎样才能知道一个模型对未曾出现过的场景仍然具有很好的泛化能力。在目前的实践中，数据集被随机划分为训练集和测试集，模型也相应地在这个数据集上被训练和评估。在这种做法中，测试集拥有和训练集一样的数据分布，因为它们都是从具有相似场景内容和成像条件的数据中采样得到的。然而，在实际应用中，测试图像或许会来自不同于训练时的数据分布。这些未曾出现过的数据可能会在视角、大小尺度、场景配置、相机属性等方面与训练数据不同。当前模型对数据分布自然变化的敏感性可能成为计算机视觉模式识别的一个严重问题。

另一个重要的挑战是如何更好地利用小规模训练数据。虽然深度学习通过利用大量标注数据在各种任务中都取得了巨大的成功，但现有的技术通常会因为只有很少的标记实例可用而在小数据情景中崩溃。这个情景通常被称为“少样本学习”。如何赋予神经网络像人类这样的泛化能力是一个开放的研究问题。另一个极端是如何利用超大规模数据有效地提高识别算法的性能。对于像自动驾驶这样的关键应用，图像识别的出错成本非常高。因此，研究者们创造出了非常庞大的数据集，这些数据集包含了数以亿计的带有丰富标注的图像，并且希望通过利用这些数据使模型的准确度得到显著提高。然而，目前的算法并不能很好地利用这种超大规模数据。在大规模数据的情况下，继续增加训练数据带来的收益会变得越来越不明显，这也是一个有待解决的重要问题。

4 与人的智能的异同

人类拥有视觉器官，光信号传入眼球在大脑皮层中形成视觉图像，即使是人类婴儿也可以轻松感知和识别图像。计算机视觉模式识别是用计算机模拟实现人的视觉功能，对客观世界的场景的进行感知和识别。两者的目标是一致的，而且研究者也是根据人类智能启发得到计算机视觉模式识别的一些经典模型，比如受到人类神经系统启发而设计的神经网络。但由于计算机与人物理结构的不同，注定了计算机只能尝试去模拟人的智能：正如人们尝试从鸟的飞行中获得启发，最终造出了飞机，诞生了空气动力学这一学科。所以，计算机视觉模式识别与人的智能也有很大的区别。人类视觉分辨率较低，而计算机可以通过高清摄像头甚至是红外摄像头来获取图像，在数据精度和来源上比人类要好得多。但是，人类可以自然地用人的智能在大脑中识别出图像中的物体并准确分类，而计算机在这方面则要困难得多。比如人类可以根据几副图像就可以辨别出什么是猫，而计算机则需要大量标注正确的数据才能将猫与其他物体分开。一个可能的解释是，人类智能有各种各样的先验知识，可以完成对事物的抽象、推理、归纳、联想和预测，使得其具有极强的泛化能力，能在细粒度上把握物体的本质特征，而计算机则犹如一张白纸：它不知道任何其他的信息，仅仅拥有输入的数据，连逻辑推理和归纳的能力都没有（当然可以有引入知识库作为先验知识并结合机器学习技术来进行逻辑推理的人工智能，但目前进展缓慢）。而且，现有的计算机视觉模式识别大多依靠深度学习技术，是一个“黑箱子”：我们并不知道计算机是如何学会的，计算机的“学习过程”对于人们来说是不可见的，人们只能看到

计算机的输出结果，而对结果的解释仍然是学界悬而未解的一个开放问题。

5 结束语

本文介绍了什么是计算机视觉模式识别，基于计算机视觉模式识别的几个重要任务介绍领域的进展，由此引入计算机视觉模式识别领域乃至人工智能领域的两大挑战，最后介绍了计算机视觉模式识别与人的智能的异同。

Reference:

[1] https://en.wikipedia.org/wiki/Computer_vision

[2] Kidiyo Kpalma, Joseph Ronsin：An Overview of Advances of Pattern Recognition Systems in Computer Vision，2007

[3] Blessing of Dimensionality: High-dimensional Feature and Its Efficient Compression for Face Verification，2013

[4] Matthew A. Turk, Alex P. Pentland：Face Recognition Using Eigenfaces，1991

[5] Renliang Weng，Jiwen Lu，Junlin Hu，Gao Yang：Robust feature set matching for partial face recognition，2013

[6] Erjin Zhou, et al：Naive-Deep face Recognition: Touching the Limit of LFW Benchmark or Not?，2015

[7] Deepface: Closing the gap to humal-level performance in face verification，2014

[8] Girshick R, Donahue J, Darrell T, et al.：Rich feature hierarchies for accurate object detection and semantic segmentation，2014

[9] Redmon J, Farhadi A：YOLO9000: Better, faster, stronger，2017

[10] Yuheng, S, Hao, Y： Image Segmentation Algorithms Overview，2017