第3章　计算机视觉技术及应用

视觉是一个生理学词汇。光作用于视觉器官，使其感受细胞兴奋，其信息经视觉神经系统加工后便产生视觉。通过视觉，人和动物感知外界物体的大小、明暗、颜色、动静，获得对机体生存具有重要意义的各种信息，至少有80%的外界信息经视觉获得，因此视觉是人和动物最重要的感觉。计算机视觉是一个跨学科的科学领域，它主要研究计算机如何从数字图像或视频中获得高水平的理解。从工程学的角度来看，它试图理解和自动化人类视觉系统可以完成的任务。计算机视觉任务包括获取、处理、分析和理解数字图像的方法以及从现实世界中提取数据以产生数字或符号信息。这种图像理解可以看作借助几何、物理、统计学和学习理论构建的模型将符号信息从图像数据中分离出来。

3.1　计算机视觉的定义与发展

3.1.1　计算机视觉的定义

顾名思义，计算机视觉是用计算机来看世界的科学。使用摄像机和计算机来代替人眼和人脑来观察分析图像和视频，对其中的目标进行识别、跟踪、测量。计算机视觉通过将图像与其中的多维数据建立起联系从而获取更多的信息。计算机视觉是一门综合性的工程学科，它包含了计算机科学、信号处理、物理学、应用数学、统计学、生物学、认知科学等多种学科。

随着智能时代的到来，计算机以及智能化产品将越来越深入地渗透到我们的生活中。计算机的功能日益强大，同时所需要的技术也更加复杂了。为了解决计算机使用起来复杂而死板的规则，使计算机能够更加便捷地被使用，需要让它来适应我们的习惯和需求，而不是我们用死记硬背的方式来使用它。计算机视觉最终的目标是让计算机像人类的大脑一样通过视觉观察和理解世界，并主动地适应环境，当然要想实现这个远大的目标还需付出巨大的努力。

3.1.2　计算机视觉的发展

计算机视觉经历了漫长的发展。从20世纪中期开始，计算机视觉经历了从二维图像到三维图像再到视频的不断探知，算法也从简单的神经网络发展到深度学习。

20世纪50年代，神经生物学家David Hubel和Torsten Wiesel在对猫的视觉实验中发现了视功能柱结构，移动边缘对视觉的初级皮层神经元有敏感刺激。对视觉神经的研究为计算机视觉奠定了基础。在同一阶段，Russell和他的同学研制了第一台数字图像扫描仪，这个仪器可以将图片转化为灰度值，从此数字图像处理迎来了开端。

20世纪60年代，Lawrence Roberts的《三维固体的机器感知》开创了以理解三维场景

为目的的计算机视觉，其中对积木的边缘、角点、线条、平面等分析给人们带来了极大的启发。1966 年，麻省理工学院人工智能实验室启动了夏季视觉项目，设计一个可将前景和背景自动分割并实现非重叠物体提取的平台，虽然没能成功地实现这个平台，但也标志着计算机视觉正式作为一个科学领域。1969 年，贝尔实验室研发出用于光子转化的电脉冲电荷耦合器件，能够应用于高质量的数字图像采集任务中。计算机视觉于这个阶段正式投入了市场应用。

20 世纪 70 年代，麻省理工学院人工智能实验室正式开设计算机视觉课程，并提出了计算机视觉理论，此理论和积木世界的分析方式有较大的不同，成为了计算机视觉下一阶段发展的重要理论框架。

20 世纪 80 年代，计算机视觉从理论走向了应用。1980 年，日本科学家 Kunihiko Fukushima 提出了一个名为 Neocognitron 的人工卷积网络。这个网络是第一个神经网络，它是卷积神经网络(Convolutional Neural Network，CNN)中卷积层和池化层的灵感来源。1982 年，David Marr 所著《视觉》一书的问世，标志着计算机视觉成为一门独立学科。同年，日本 COGEX 公司生产了世界第一套工业光学字符识别系统 Dataman。1989 年，法国的 Yann LeCun 在 Neocognitron 的基础上应用了一种后向传播风格，并在几年后发布了 LeNet-5 网络。卷积神经网络已经成为图像识别中的重要组成部分。

20 世纪 90 年代，人们开始致力于研究特征识别。1997 年，伯克利教授 Jitendra Malik 发表一篇论文，试图使用机器对图像进行自动分割。1999 年，David Lowe 发表了《基于局部尺度不变特征的物体识别》一文。同年，Nvidia 公司提出了图形处理单元(Graphic Processing Unit，GPU)的概念，用于为执行复杂的属性计算而设计的数据处理芯片，GPU 的到来为多种行业的发展提供了动力。

21 世纪初，计算机视觉的发展走向了高潮。2001 年，Paul Viola 和 Michael Jones 研发了第一个可以实时工作的人脸检测框架。2005 年，方向梯度直方图的方法被提出并应用于行人检测，是计算机视觉、模式识别很常用的一种特征检测方法。2006 年，空间金字塔算法提出并用于进行图像匹配、识别和分类。同年，PASCAL VOC 提供了开源的数据库以及对数据进行注释的工具，并举办了年度竞赛，使得更多的研究人员加入图像识别的研究中来。同年，Geoffrey Hilton 提出了深度置信网络，并为多层神经网络赋予了深度学习这个新名字。2009 年，可变形零件模型(Deformable Parts Model，DPM)算法诞生，它是在深度学习大范围发展以前最好的目标识别算法，此算法在行人检测任务中达到了十分优异的效果，研究出这个算法的 Felzenszwalb 教授也被 VOC 授予终身成就奖。

21 世纪 10 年代，深度学习在计算机视觉中被广泛使用。2009 年，李飞飞教授发布了一篇名为 *ImageNet：A Large-Scale Hierarchical Image Database* 的论文，并发布了 ImageNet 数据集，此数据集从 2010 年到 2017 年共参与了 7 届 ImageNet 挑战赛。它改变了人们对数据集的认识，发现了数据集和算法一样重要，推动了计算机视觉和深度学习的发展。2012 年，Alex Krizhevsky 创造了 AlexNet，它是第一个在 ImageNet 数据集上表现极为出色的算法，它使机器识别的错误率从 25%下降到 16%，真正地展示了 CNN 的优点。2014 年对抗网络诞生，是计算机视觉领域的一大突破。2016 年 Facebook 的 DeepFace 人脸识别算法达到了 97.35%的准确率，几乎与人眼不分上下。2017 年，特征金字塔网络提出，可以从图像中提取出更加深层的语义信息。随着计算机视觉和深度学习的紧密结合以及计

算机算力的不断发展,各种视觉任务达到了更好的完成结果。计算机视觉更多地进入了人们的实际生活应用中。

3.1.3 计算机视觉相关学科

有许多学科都与计算机视觉的知识十分相似,像数字图像处理、模式识别、图像理解等。它们之间联系紧密,计算机视觉是在图像处理和模式识别等相关学科的基础上发展来的,它的最终目标是实现图像理解。

1) 图像处理

图像处理技术是指将图像用计算机进行分析,转化为另一幅包含更多特征的图像。常用的图像处理技术包括图像压缩、增强复原、匹配和识别等。图像处理后的结果可以作为下一步图像结果的分析,也可以作为处理的最终结果,计算机视觉中常用图像处理作为特征提取的手段。

2) 模式识别

模式识别是指用计算机根据不同图像中特征的不同,将图像划分为不同的类别。模式识别也可以称为模式分类,可以分为有监督的分类和无监督的分类,模式还可以分为抽象和具体两种形式。模式识别主要研究生物如何感知物体以及如何在给定的任务下用计算机实现模式识别这两个方面。模式识别与统计学、心理学、语言学、计算机科学等多种学科都有联系,与图像处理、计算机视觉等研究有交叉关系。

3) 图像理解

图像理解指的是给计算机一张图像,计算机不但能描述图像本身,还可以对图像内的物体做出解释,研究图像中有哪些目标,目标之间有什么样的关联,图像所处的场景是怎样的。图像理解以计算机视觉为载体来模拟人类视觉,是计算机视觉的最终目的。

计算机视觉所涉及的学科众多,上述的几种学科以及很多其他的学科都有着密切的关系,因此计算机视觉是一个极为复杂、研究领域极广的学科。

3.2 深度学习与计算机视觉

3.2.1 深度学习

深度学习是机器学习中的一个领域,它是通过对数据集或样本库进行深层次的理解与学习,对图像、视频、文字、声音等多个数据进行研究。深度学习在搜索技术、机器翻译、计算机视觉、自然语言处理、个性化推荐等多个领域都发挥了极大的作用。

深度学习从研究内容来看可以分为三类,分别是基于卷积计算的神经网络系统(常称为卷积神经网络)、基于多层神经元的自编码神经网络和深度置信网络。随着对深度学习研究的深入,科研人员逐渐将不同的方法和不同的训练步骤相结合,以达到更加优秀的训练结果。与传统的方法相比较,深度学习中设置了更多的参数模型,因此参与训练的数据量更大,模型的训练难度更大,但训练达到的效果会更好。

深度学习有以下几个优点。

(1) 学习能力强。

(2) 覆盖范围广,有较强的适应性,可以解决复杂问题。

(3) 数据量越大,表现效果越好。

(4) 多平台多框架兼容。

深度学习也存在以下缺点。

(1) 由于所需算力和数据规模过大,难以在移动设备上使用。

(2) 对硬件的要求高。

(3) 使用困难,模型设计复杂。

(4) 过于依赖数据,可解释性不高,当数据种类不平均时会产生较大误差。

深度学习的本质是人工神经网络,深度神经网络指的是具有一层及一层以上的隐含层的神经网络,通常用于对复杂的非线性系统进行建模,其中常用的几种网络结构如下。

1) CNN

CNN 是为了完成生物视知觉仿造任务而构造的,是一种包含卷积计算且具备深度结果的前馈神经网络,可以用监督学习和非监督学习进行训练。CNN 可以对数据进行平移不变的分类,因此也称为平移不变人工神经网络。CNN 的网络架构如图 3-1 所示。

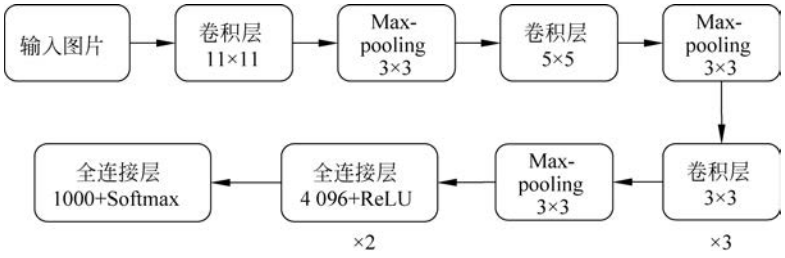

图 3-1 CNN 网络架构

2) 深度信念网络

深度信念网络(Deep Belief Network,DBN)是一种包含多层隐藏层的概率生成模型,与传统的神经网络判别模型相对比,生成模型对数据和标签进行联合对比观察。DBN 由多个限制玻尔兹曼机层构成,采用无监督逐层训练的方式进行训练,可以对训练的数据进行深层次的表达。DBN 网络架构如图 3-2 所示。

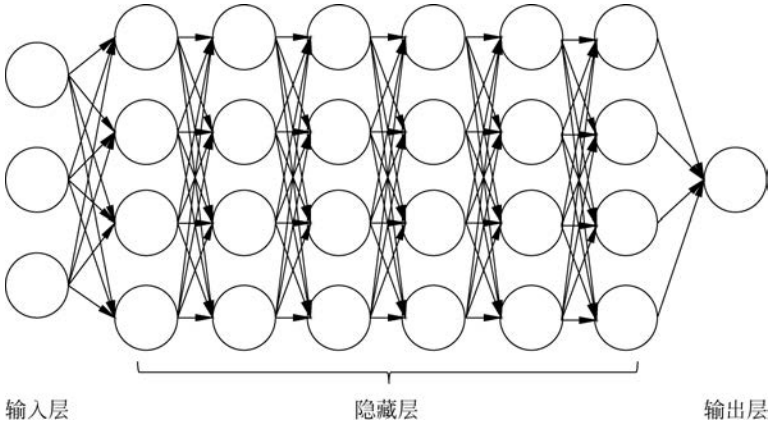

图 3-2 DBN 网络架构

3) 循环神经网络

循环神经网络(Recurrent Neural Network,RNN)是以序列数据为输入并在序列的方向上进行递归的递归式神经网络,网络内的循环单元按链式相连接。RNN 由于其记忆性的特点,在对序列数据进行学习时有一定的优势,常被应用在各种时间序列预测中。CNN 和 RNN 相结合的神经网络可以用来处理输入为序列的计算机视觉问题。RNN 网络架构如图 3-3 所示。

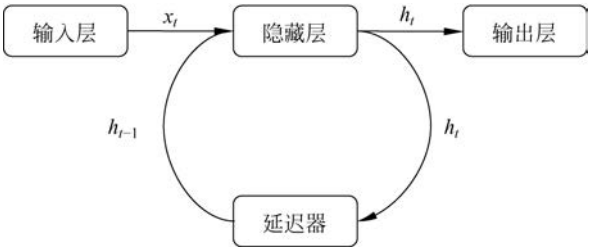

图 3-3 RNN 网络架构

4) 监督学习

监督学习是指参与训练的数据都带标签,且训练的误差是从上向下传输的训练过程。监督学习的第一步是对输入数据进行学习,得到各层的参数并进一步对多层模型的参数进行优化调整。监督学习第一步得到的初始值接近全局最优,因此取得的效果更好。

5) 无监督学习

无监督学习是指参与训练的数据不带标签,从底层开始一层一层向上的训练过程。由于人工给数据进行分类打标签的任务成本过高,因此需要计算机来帮助实现这一目标。首先用没有标签的数据训练第一层并学习到数据本身的结构,得到比输入的数据更加有表现能力的输出,并输入下一层中。学习到 $n-1$ 层时,将输出作为 n 层的输入,从而做到自下而上地训练,并得到各层的参数。

3.2.2 深度学习在计算机视觉中的应用

传统的视觉算法通常包含 5 个步骤,分别为特征感知、图像预处理、特征提取、特征筛选和推理预测与识别,并且传统的特征提取主要依靠人工完成,对于简单的任务来说效果好,但对于规模较大的数据集难以实现。

深度学习在处理信息量较为丰富的任务上有很好的表现,非常适合计算机视觉任务,大规模的数据集和深度学习网络的强大能力为计算机视觉提供了广阔的发展空间。随着深度学习的加入,计算机视觉从最初的图像变换、图像编码压缩、图像增强与复原、图像分割、图像描述等逐渐扩散到更加复杂的领域,生活中最为常见的图像分类、识别应用有人脸识别、指纹识别、车牌识别等。

1) 局部卷积神经网络

局部卷积神经网络(Region-CNN,R-CNN),是第一个将深度学习运用到目标检测上的算法,R-CNN 的目标检测准确度与其之前的算法相比较有了大幅度的提升,原作者在 PASCAL VOC 2012 数据集上进行测试,平均准确率(Mean Average Precision,MAP)为 53.7%,相较于 DPM 算法的 35.1%提升了 18.6%。R-CNN 带来的成功让大家看到了

CNN 在计算机视觉领域上的无限潜能，越来越多的研究人员将 CNN 运用到目标检测的模型中去。

传统的目标检测一般先在图片上圈出所有可能是目标物体的区域框，然后对这些区域框进行特征提取并使用图像识别的方法分类，分类后的区域用非极大值抑制的方法进行输出。R-CNN 保存着传统的目标检测的思路，保留使用区域框进行特征提取、图像分类、非极大值抑制的方法，区别在于将传统的特征提取方法换成了深度卷积网络特征提取的方法。

R-CNN 的具体步骤如图 3-4 所示，可以分为以下几个步骤。

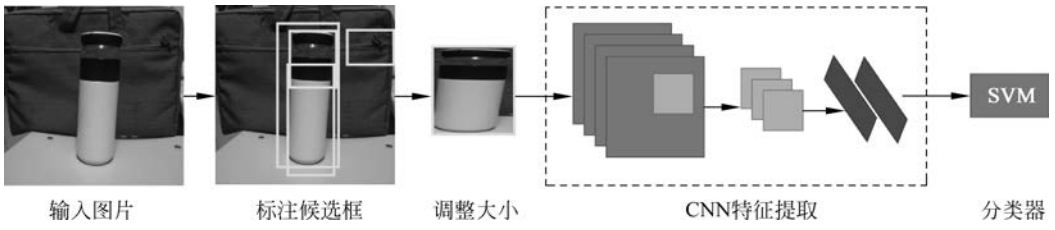

图 3-4　R-CNN 具体步骤

① 对输入的一张图片采用选择性搜索(Selective Search，SS)算法提取 2000 个类别独立的区域框。

② 将每个区域框调整为固定的大小，用 CNN 提取特征向量。

③ 对每个区域框进行支持向量机(Support Vector Machine，SVM)目标分类。

④ 训练一个边界框回归模型，对框的准确位置进行修正。

2) 常见数据集

数据集是深度学习中不可缺少的部分，深度学习的学习都是基于数据集内大量数据所携带的信息，训练用的数据集量越大，得到的训练结果可能会越好。计算机视觉所需要的数据集比较庞大，且个人收集起来十分复杂，因此网络上有许多公开的数据集可供研究人员学习使用。

下面列举几个常用的开源数据集。

① ImageNet：该数据库根据 WorldNet 层次结构进行组织，WorldNet 中有超过 100 000 个同义词集，其中 ImageNet 为每个同义词集提供 1000 个左右的图像进行说明。

② MS COCO：数据集包含 91 个对象类型的照片，照片中的目标清晰、易于识别，在 3 283 000 个图像中共有 2 500 000 个带标签的实例。

③ Cityscapes：用于做城市街景理解的数据集，数据集分为测试集和验证集，测试集无标注，包含 50 个城市的不同场景、不同背景、不同时间段的街景图片，其中 5000 个为精细标注，20 000 个为粗略标注。

④ KITTI：自动驾驶领域使用最广泛的数据集之一，可用于评测立体图像、光流、视觉测距、三维物体检测等任务，数据采集平台包括 2 个灰度摄像机、2 个彩色摄像机、一个 Velodyne 3D 激光雷达、4 个光学镜头以及 1 个 GPS 导航系统。一共细分为道路、城市、住宅、校园和人 5 类数据；包含市区、乡村高速公路的数据，每张图像最多 15 辆车及 30 个行人，而且还包含不同程度的遮挡。整个数据集由 389 对立体图像和光流图、39.2 千米视觉测距序列以及超过 200 000 个 3D 标注物体的图像组成。

3.3 计算机视觉关键技术

3.3.1 特征检测

在计算机视觉技术中,特征检测是十分基础而重要的技术。计算机视觉中的多种任务,如目标识别、图像分类、图像分割、立体视觉、三维重建等工作都是以特征检测为基础的,通过对特征的检测与提取从而完成后续任务。特征检测中的特征包括特征点、轮廓、边缘等,有明显的可以识别的与周围环境差异较大位置都是特征。

生活中随手拍摄的照片都可以用于特征检测,图 3-5(a)是一张手机拍摄的风景图片,图 3-5(b)是从图 3-5(a)中截取出的校徽部分并进行了放大。用眼睛可以轻易地分辨识别出图 3-5(b)是图 3-5(a)的哪一部分,而计算机视觉技术则是通过检测两张图像中的特征点,判断相同的特征点来进行匹配。特征点也可以称为兴趣点、角点,是图像的重要因素之一,指的是图像中关键、显而易见的点,如图像中某个部分的边角点、特殊形状物体的边缘端点等。经过特征检测后,图 3-5(a)中图像的特征点用圆圈圈出来,如图 3-6 所示,图片中的字、建筑物的边角点、树枝的末端、校徽内不同颜色的交界点等都是特征点。

(a) 原始图片　　(b) 截取部分图片

图 3-5　截取部分图片用于特征检测示意图

图 3-6　特征点检测

特殊点可以用来寻找不同图像中特殊点相同的对应部分,下面通过特殊点的识别与匹配将图 3-5 中的两张图片匹配起来,如图 3-7 所示,可以看出两张图片中相同的特殊点用直线相连接,通过检测两张图像的特殊点,并对特殊点进行比对,相同特征的点即可对应连接

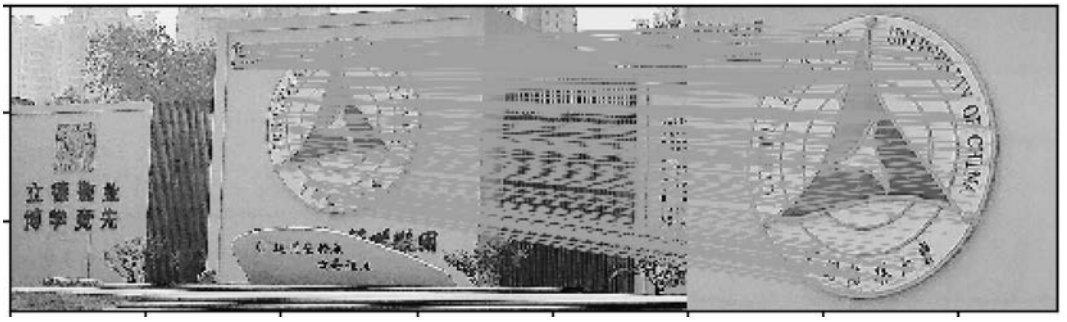

图 3-7　特征点匹配

匹配起来。

用于进行特殊点检测的算子称为特征描述算子，常用的特征描述算子有尺度不变特征检测、Harris 特征点检测、偏差和增益规范化检测等，下面重点介绍尺度不变特征检测。

尺度不变特征变换(Scale-Invariant Feature Transform，SIFT)是一种经典的局部特征描述算子，最初是由 David Lowe 于 1999 年发表的。

SIFT 的原理是将图像高斯模糊后，图像中不同区域点的变化不同，变化较小的点一般为平滑区域，变化较大的点则为特征点。将检测到的关键特征点作为中心，选择 16×16 的窗口，将这个区域平均分为多个 4×4 的子区域，每个 4×4 子区域分成 8 个区间，即可得到 4×4×8=128 维度的特征向量。SIFT 算法主要可以分为 4 个步骤。

1) 尺度空间极值检测

尺度空间是在进行图像处理的模型内引入一个尺度的参数来使其拥有尺度不变性的特征，通过对空间内的各个尺度的图像进行处理，模拟人眼距离看到目标的远近差异的过程，对图像进行逐渐增长的模糊处理，图像的模糊程度与尺度成正比。用图像和高斯函数进行卷积得到图像的高斯尺度，如式(3-1)所示。

$$L(x,y,\sigma)=G(x,y,\sigma)*I(x,y) \tag{3-1}$$

$$G(x,y,\sigma)=\frac{1}{2\pi\sigma^2}e^{\frac{x^2+y^2}{2\sigma^2}} \tag{3-2}$$

其中 $L(x,y,\sigma)$是高斯尺度，σ 是尺度空间因子。

上文描述了尺度空间的定义，接下来通过高斯金字塔的方式来实现尺度空间的搭建。高斯金字塔是通过将图像逐层高斯滤波并进行降阶采样，得到的图像进行由大到小排列构成金字塔状，金字塔模型的最底下一层为原始的图像。首先对原始图像进行不同参数的高斯滤波，得到多张模糊程度不同的图像，然后进行降阶采样后得到上一层的图像，得到的图像作为再上一层的原始图像，重复进行操作直到满足层数需求。金字塔每层的图像进行多参数高斯模糊，因此塔每层都包含多张图像，每层的多张图像组合称为 Octave，这些图像的大小一致但模糊程度不同。

在 SIFT 特征点检测中选择了差分高斯金字塔代替高斯金字塔，可以有效地提高检测的效率。如图 3-8 所示为尺度空间极值检测中的工作流程。

在尺度空间内寻找极值点，每个监测点需要与以其为中心的周围 3×3 范围内的 8 像素

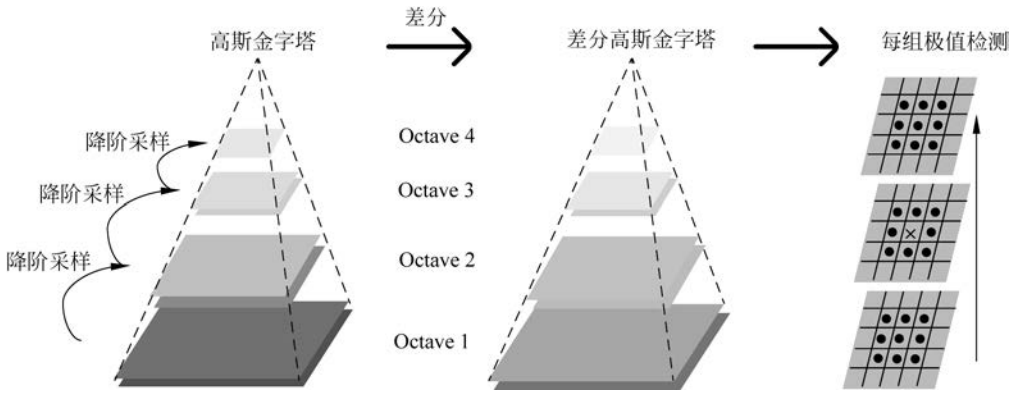

图 3-8　尺度空间极值检测

以及上下两层 3×3 区域内的 9×2 像素(共 26 像素)相比较,当它大于或小于相邻的像素时,则为极值点。

2) 精确特征点的位置

由于数字图像都为离散采样的图像,而实际的图像是连续的,并且还需要考虑在边缘位置的极值点,因此在上一步骤中检测出的极值点有可能出现偏差。因此要对差分高斯空间进行拟合处理,来精确特征点的位置。

通过设置阈值来判断极值点是否在边缘上,$\boldsymbol{H}(x,y)$为差分高斯金字塔中对 x 和 y 的二阶导数。

$$\boldsymbol{H}(x,y)=\begin{bmatrix}D_{xx} & D_{xy}\\ D_{xy} & D_{yy}\end{bmatrix} \tag{3-3}$$

$\mathrm{Tr}(\boldsymbol{H})$为矩阵 $\boldsymbol{H}$ 的迹,$\mathrm{Det}(\boldsymbol{H})$是行列式。

$$\mathrm{Tr}(\boldsymbol{H})=D_{xx}+D_{yy}=\alpha+\beta \tag{3-4}$$

$$\mathrm{Det}(\boldsymbol{H})=D_{xx}D_{yy}-(D_{xy})^2=\alpha\beta \tag{3-5}$$

若极值点不满足下式,则舍去该点。

$$\frac{\mathrm{Tr}(\boldsymbol{H})^2}{\mathrm{Det}(\boldsymbol{H})}<\frac{(r+1)^2}{r} \tag{3-6}$$

3) 确定特征点的方向

通过对图像的每个关键点赋予一个方向,可以使得这个特征检测算子具有旋转不变性,也就是当目标发生方向的变化时,只要其他的特征都相对应,也可以识别出。

极值点的方向通过其周围的像素的梯度来确定,梯度的公式如下:

$$\mathrm{grad}(I(x,y))=\left(\frac{\partial I}{\partial x},\frac{\partial I}{\partial y}\right) \tag{3-7}$$

梯度的幅值为

$$m(x,y)=\sqrt{(L(x+1,y)-L(x-1,y))^2+(L(x,y+1)-L(x,y-1))^2} \tag{3-8}$$

梯度的方向为

$$\theta(x,y)=\arctan\left(\frac{L(x,y+1)-L(x,y-1)}{L(x+1,y)-L(x-1,y)}\right) \tag{3-9}$$

用直方图来对特征点的方向进行统计，将 0 度到 360 度分为 36 个部分，每个部分表示 10 度，只要大于最大峰值 80％则认为是该特征点的辅助方向。

4）特征点的描述

经过上述步骤产生的特征点都是基于图片的点坐标的，如果想根据特征点与其他的图像进行对比，需要将特征点单独提取出来。通过对特征点周围进行分块，并计算梯度直方图，生成具有唯一性的方向向量来代表这部分的图像，从而产生 SIFT 特征向量。

3.3.2 图像分割

图像分割，顾名思义即为将想要识别的目标从图像中分割出来。图像分割是计算机视觉中十分重要的任务，它在实际生活中有广泛的应用，并发挥着核心的作用，例如，在行人检测、视频监控、自动驾驶、医学图像分析等方面，图像分割都扮演着不可或缺的角色。图 3-9 为图像分割的例子。

图 3-9 图像分割

图像分割可以分为两大类，一个是语义分割，另一个是实例分割。

1）语义分割

语义分割指的是将图像中的待识别目标分割出来，并对分割的目标进行分类。图像中一般都会同时存在多种物体，语义分割根据像素级别将图片分为多个部分，分割出不同类别的目标。

2）实例分割

实例分割指的是将图像中的待识别目标分割出来，对分割的目标分类后，还需要对分类后的目标进行区分，将每个不同的实例单独分割。相较于语义分割，实例分割将每一个目标作为一个待分割的实例。

实例分割并不是一个独立的任务，它是通过语义分割发展演变而来的。从最初的简单的算法开始，图像分割算法经过了多年的改变与进化，达到了越来越好的分割效果。图像分割部分算法发展历史如图 3-10 所示。

图像分割算法可以按照分割方式的不同分为以下 5 种。

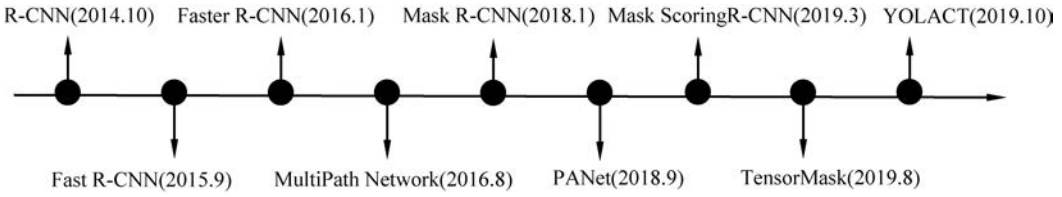

图 3-10　图像分割算法的发展

① 阈值分割方法。选取一个合适的像素值作为边界,将图像处理成对比度较高的、分割部分容易识别的方法。

② 区域增长细分方法。通过将属性相似的像素组合在一起形成一个区域,在区域内找到一个种子像素,将周围的属性与种子像素相似的像素合并到区域中。将这些新合并进来的像素作为新的种子像素继续合并,可以得到没有满足属性相似要求的像素。

③ 边缘检测分割方法。该方法主要通过图像的灰度值不同以及边缘的突出进行分割。

④ 基于聚类分割方法。通过将类的划分以物体间的相似性作为标准,使相似的类别尽可能相似,不相似的类别区别尽可能大。

⑤ 基于 CNN 的弱监督学习分割方法。对图像内待识别对象区域用部分像素进行标记。基于 CNN 的分割算法是图像分割任务中的研究热点,图 3-10 展示了基于 CNN 的图像分割的重要算法自 2014 年到 2019 年的发展。

3.3.3　R-CNN 系列算法

上文中已经介绍过 R-CNN 算法的细节,接下来介绍基于 R-CNN 的几种算法的演进。

Fast R-CNN 在 R-CNN 的基础上进行了一些变动,即在 R-CNN 的最后一个卷积层后添加感兴趣区域(Regions of Interest,ROI)的池化层。与 R-CNN 先提取特征值,然后 CNN 提取特征放入 SVM 分类器,之后做 bbox 回归的步骤不同,Fast R-CNN 将 bbox 回归与区域在神经网络内部合并成为多重任务模型,并使用 Softmax 代替了 SVM 分类器。Fast R-CNN 的改进有效地解决了 R-CNN 严重的速度问题,并且为 Faster R-CNN 做了铺垫。

Faster R-CNN 在 Fast R-CNN 的基础上使用了区域生成网络(Region Proposal Network,RPN)来生成候选框,让 RPN 和 Fast R-CNN 共享 CNN 特征,成为一个端到端的 CNN 对象检测模型。

Mask R-CNN 算法在 Faster R-CNN 的基础上创新了 ROI 对齐操作,引用全卷积网络(Fully Convolutional Network,FCN)生成 Mask,并且添加了用于语义分割的 Mask 损失函数,改变了算法损失函数的计算方法。

Mask Scoring R-CNN 解决了 Mask R-CNN 的一个重要的问题。Mask R-CNN 中使用了边框的分类置信度作为 Mask 的分数,但通过边框的分类置信度作为 Mask 准确率是不准确的,会导致预测结果相差较大。因此 Mask Scoring R-CNN 创新出了一种新方法,添加 MaskIoU Head 模块,将 Mask Head 操作后得到的预测分数与 ROI 特征输入卷积层和全连接层,从而得到模型的分数。

表 3-1 所示为这几种算法的对比。

表 3-1　几种基于 R-CNN 的算法对比

算法名称	使用方法	缺点	改进
R-CNN	① 选择性搜索 SS 提取候选区域(Region Proposal,RP)。 ② CNN 提取特征。 ③ SVM 分类。 ④ bbox 回归	① 训练步骤烦琐。 ② 训练所占空间大。 ③ 训练耗时长	MAP 为 66%
Fast R-CNN	① SS 提取 RP。 ② CNN 提取特征。 ③ Softmax 分类。 ④ 多任务损失函数边框回归	没有实现端到端训练测试	MAP 提升至 70%； 测试耗时缩短
Faster R-CNN	① RPN 提取 RP。 ② CNN 提取特征。 ③ Softmax 分类。 ④ 多任务损失边框回归	计算量依旧比较大	测试精度和速度提升；实现端到端目标检测；迅速生成建议框
Mask R-CNN	① RPN 提取 RP。 ② ResNet-FPN 提取特征。 ③ ROI 对齐的方法来取代 ROI 池化。 ④ Mask 分支	边框分类置信度用来作为 Mask 准确率时不够精确	ROI 对齐能将像素对齐，满足了图像语义分割的准确度要求
Mask Scoring R-CNN	① RPN 提取 RP。 ② ResNet-FPN 提取特征。 ③ 加入 MaskIoU 分支		获得更加可靠的 Mask 分数

图像分割评分指标有很多，如下所示。

(1) 平均正确率(Average Precision,AP)，指的是所有类别的正确率。

$$\mathrm{AP}=\int_0^1 p(r)\mathrm{d}r \tag{3-10}$$

(2) 像素精度(Pixel Accuracy,PA)，指标记正确的像素占全部像素的比例。

$$\mathrm{PA}=\frac{\sum_{i=0}^{k} p_{ii}}{\sum_{i=0}^{k}\sum_{j=0}^{k} p_{ij}} \tag{3-11}$$

(3) 均像素精度(Mean Pixel Accuracy,MPA)，指在 PA 的基础上对标记正确像素占全部像素的比例做类平均。

$$\mathrm{MPA}=\frac{1}{k+1}\sum_{i=0}^{k}\frac{p_{ii}}{\sum_{j=0}^{k} p_{ij}} \tag{3-12}$$

(4) 交并比(Intersection over Union,IoU)，指计算真实值和预测值两个集合的交集与并集之比。

$$\mathrm{IoU}=\frac{\mathrm{Area\ of\ Overlap}}{\mathrm{Area\ of\ Union}}=\frac{A_{\mathrm{pred}}\cap A_{\mathrm{true}}}{A_{\mathrm{pred}}\cup A_{\mathrm{true}}}=\sum_{i=0}^{k}\frac{p_{ii}}{\sum_{j=0}^{k} p_{ij}+\sum_{j=0}^{k} p_{ji}-p_{ii}} \tag{3-13}$$

(5) 均交并比(Mean Intersection over Union,MIoU),指在每一类上计算 IoU 后进行平均。MIoU 是使用最频繁的图像分割精准度度量标准。

$$\mathrm{MIoU}=\frac{1}{k+1}\sum_{i=0}^{k}\frac{p_{ii}}{\sum_{j=0}^{k}p_{ij}+\sum_{j=0}^{k}p_{ji}-p_{ii}} \tag{3-14}$$

(6) 频权交并比(Frequency Weighted Intersection over Union,FWIoU),指在 MIoU 的基础上进行升级,根据类别出现的频率设置权重。

$$\mathrm{FWIoU}=\frac{1}{\sum_{i=0}^{k}\sum_{j=0}^{k}p_{ij}}\sum_{i=0}^{k}\frac{p_{ii}}{\sum_{j=0}^{k}p_{ij}+\sum_{j=0}^{k}p_{ji}-p_{ii}} \tag{3-15}$$

3.3.4 立体视觉

立体视觉指的是用两个或多个摄像头来获取深度的视觉信息的技术。

首先介绍双目视觉求解深度。双目视觉求解深度就是根据透视几何图形学的三角化原理,通过左边拍摄的图像上面的任意一个点,在右边拍摄的图像上找到相应的匹配点,即可确定该点的三维坐标。图 3-11 所示为双目视觉求深度的过程。

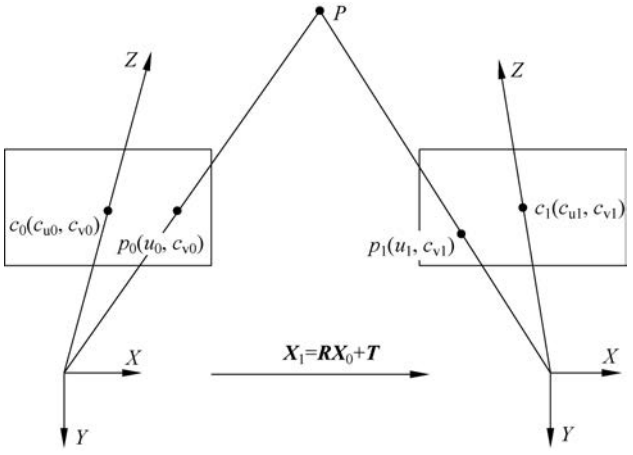

图 3-11 双目视觉求深度过程

在图 3-11 中,P 点为选中的任意一点,P 点在左右两个相机中成像的位置分别为 p_0 和 p_1,两个相机的焦距分别为 f_0 和 f_1,且两个相机的相对位移分别为 $\boldsymbol{R}$ 和 $\boldsymbol{T}$。根据小孔成像原理可知

$$\boldsymbol{z}_0\begin{bmatrix}u_0\\v_0\\1\end{bmatrix}=\begin{bmatrix}f_{x0}&0&c_{u0}\\0&f_{y0}&c_{v0}\\0&0&1\end{bmatrix}\begin{bmatrix}x_0\\y_0\\z_0\end{bmatrix} \tag{3-16}$$

$$\boldsymbol{z}_1\begin{bmatrix}u_1\\v_1\\1\end{bmatrix}=\begin{bmatrix}f_{x1}&0&c_{u1}\\0&f_{y1}&c_{v1}\\0&0&1\end{bmatrix}\begin{bmatrix}x_1\\y_1\\z_1\end{bmatrix} \tag{3-17}$$

由相机的左右相对位置关系 $\boldsymbol{X}_1=\boldsymbol{R}\boldsymbol{X}_0+\boldsymbol{T}$ 可得

$$\begin{bmatrix} x_1 \\ y_1 \\ z_1 \end{bmatrix} = \begin{bmatrix} r_{00} & r_{01} & r_{02} \\ r_{10} & r_{11} & r_{12} \\ r_{20} & r_{21} & r_{22} \end{bmatrix} \begin{bmatrix} x_0 \\ y_0 \\ z_0 \end{bmatrix} + \begin{bmatrix} t_0 \\ t_1 \\ t_2 \end{bmatrix} \tag{3-18}$$

因此只要找到左图上一点在右图上的匹配点，即可求出该点在相机坐标系中的坐标。那么接下来解决从右图找左图对应点坐标的问题。

一般来说，从右图中找左图中已知的对应点是一个复杂度较高的二维搜索问题，为了降低算法的复杂度，使用极线约束将此问题转换为一维问题。左图上的点在右图中可能的投影是在某一条线上，将搜索范围由面降低到线。将左右摄像头完美对齐，使它们的焦距等参数完全一致，即可将左右摄像头的极线矫正成行相同的平行线。因此左图中任意一点在右图中只能映射到与其对应的相同行上。

立体视觉的研究主要由以下几方面组成。

(1) 图像获取：立体视觉研究中需要从图像中获取许多要素，包括场景领域、时间、成像形态、分辨率、视野、摄像机的相对位置等，且图像的场景复杂度受到遮掩、人工物体、纹理区域、重复结构的区域等因素的影响。

(2) 摄像机模型：对立体摄像机组的重要几何和物理特征的表示，提供图像上对应点空间和实际场景空间之间的映射关系，还约束寻找对应点时的搜索空间。

(3) 特征抽取：特征抽取的过程即为提取匹配基元的过程。

(4) 图像匹配：是立体视觉的核心，建立图像之间的对应关系，从而计算视差。

(5) 深度计算：解决匹配问题的复杂化，提高深度计算精度。提高深度计算精度有三种方法：半像素精度估计、加长基线长、几幅图的统计平均。

(6) 内插：基于特征匹配算法得到的深度图是稀疏且分布不均匀的，而立体视觉任务中的深度图都需要是稠密的，因此基于相关区域匹配的算法更为合适，但这类算法在灰度均匀的区域匹配不准确，所以需要内插过程来近似连续深度图。

3.4 计算机视觉的实际应用

随着人工智能技术的迅速发展，人们生活越来越智能化，计算机视觉的技术也深入生活中。现如今的生活已经与十年前大相径庭，随处可见的科技化、智能化极大地方便了人们的生活。人工智能已经不知不觉中渗透进生活的每个细枝末节。

人工智能最开始受到大家的广泛关注是在人机围棋大战。2016 年 3 月，谷歌智能围棋机器人阿尔法狗以 4 比 1 的成绩战胜人类围棋世界冠军李世石，这一新闻引起了全世界的广泛关注。从此人工智能的浪花被激起，越来越多的科研人员投入这项热门科学的研究中去。计算机视觉作为人工智能的一个重要的、实用性极强的分支，更是受到极大部分研究人员的青睐。

人的生活中离不开眼睛，醒着的每分每秒都需要眼睛工作。生活中的许多工作也都是基于人眼的观察才可以完成。但人眼观察受到的限制比较多，人的记忆力、人的疲劳度都会很大程度地影响工作，并且人力劳动需要消耗费用较大。而计算机视觉正是用计算机代替人眼工作的，并且计算机的算力、速度远远强于人类，且成本较低，因此计算机视觉在生活中的实际应用十分广泛。

例如,停车场内的智能车牌识别系统、上班打卡的虹膜识别和指纹识别系统、手机应用软件中的智能物体识别功能、人脸面部表情识别、人类肢体动作识别、手写字体识别等都是生活中与人们息息相关的技术。下面详细介绍人脸识别、三维重建以及自动驾驶这三个实际应用的计算机视觉技术。

3.4.1 人脸识别

人脸识别是计算机视觉在实际应用中使用范围比较广的一项技术,在许多的场景都能见到它的身影。图 3-12 中为学生进出图书馆时,需要进行人脸识别,检测是否为本学校的学生,在很多高校的校门口和宿舍门口也设有同样的人脸识别机器。

图 3-12 学生进出图书馆进行人脸识别

随着人脸识别技术的不断提高,已经有越来越多的高级别任务开始使用人脸识别技术。例如,过去进火车站时,只需要出示身份证,检票员粗略地观察持证人与身份证上的照片是否一致,但身份证上的证件照一般为素颜且拍摄年限较长,通过人眼进行识别难免会出错,并且需要的人力成本较大。现在的火车站进站口设有多台人脸识别机器,乘车人刷身份证件同时进行面部比对,比对通过才可以顺利进站,不仅极大地节约了人力,还降低了偷用他人身份证件进站乘车的可能性。

自 2020 年新冠疫情暴发以来,为了对中国国内的疫情进行良好控制,要求大陆常住人口及入境人员出示健康码绿码才能正常进出公共场合,确保持码人员 14 天之内没有到达过发生疫情的中、高风险地区。健康码也需要通过身份证或护照与人脸进行核对方可正常显示。

还有支付宝也已经推出了人脸识别支付的方法,说明人脸识别技术的准确度已经十分高,能够确保不会错误地识别。

人脸的特征和虹膜、指纹一样,有着唯一性、不易变性以及不可复制性,因此为人的身份鉴定打下了基础。人脸识别可以分为以下主要步骤。

(1) 人脸图像的采集。

人脸识别所需要的图像即为人五官清晰的脸部图像,可以通过视频、动图、图片等多种途径获取。

(2) 人脸图像的预处理。

采集得到的包含人脸的图像不能直接用于人脸识别,需要进行预处理操作,需要对图片进行灰度变换、过滤噪声、锐化以及归一化等多种处理。

(3) 人脸特征的提取。

人脸特征的提取可以看作对图像进行关键点定位,通过图像中人的五官的位置来判断人脸的位置和大小。人脸识别就是将人脸上有用的特征信息挑出来实现人脸识别,如直方图特征、结构特征等。人脸特征提取的方法可以分为基于知识的表征方法和基于代数特征统计学的表征方法两种。比较常用的即为基于知识的表征方法,它是通过人脸的眼睛、鼻

子、嘴巴等特征点的位置结构计算它们之间距离、角度等关系，通过这些结构关系作为人脸识别的重要特征。

(4) 人脸特征的比对与匹配。

将待识别的人脸特征与数据库内的人脸特征进行搜索匹配，当特征的相似度到达一个设定的值时，即认为两者有较大的相似度，从而实现人脸识别任务。

人脸识别任务的实现中有一个部分是必不可少的，那就是数据库。数据库在人脸识别的任务中发挥了十分重要的作用。网络上可以搜集到许多公开的人脸识别数据集供大家进行科研使用，但人脸识别的应用五花八门，许多商用的人脸识别技术的数据库存在很大的安全隐私问题。

用于商用的人脸识别技术需要单独建立数据库，而数据库的建立不可避免地涉及用户的个人信息。因此数据库的安全、信息保密是十分重要的，但许多科技公司的技术和财力难以实现对用户人脸信息的保护，导致了网络上经常会出现人脸信息的售卖。

人脸识别以及其他生物识别技术的商用都给人们的生活带来了许多的便捷，但是生物信息属于不可更改的高敏感个人隐私，这些重要的数据在传输、使用、保存的过程中有极大的安全隐患问题。这些数据在未经本人知晓同意的情况下被过度分析、滥用，都严重地侵犯了个人隐私权，信息一旦泄露还会使个人行踪等更加重要的私密信息泄露。

2020 年 11 月 1 日，国家标准《信息安全技术 远程人脸识别系统技术要求》正式实施，此标准对我国人脸识别技术体系和应用场景都做出了进一步的详细约束。但对人脸等生物信息的规范管理以及生物特征识别技术的应用范围仍然需要出台更加严格的法律法规来约束，从而达到对用户隐私权和个人人身安全的良好保障。

3.4.2 三维重建

计算机视觉中的三维重建就是通过对图像进行处理，分析图像中隐含的信息来重建图像所处的三维环境。三维重建技术是环境感知的重要技术之一，自动驾驶、虚拟现实技术、增强现实技术、运动目标检测、行为分析等多种计算机视觉的实际应用中都存在着三维重建的身影。三维重建是计算机视觉中重要的部分，目标识别的任务只是计算机视觉中的比较浅层面的技术，人的视觉能够真切地感知到三维的世界，因此计算机视觉最终也会在识别的基础上走向三维的世界。

三维重建一般是通过单一的视图或者多角度的视图来对当前环境进行三维信息还原的过程。多角度的视图所包含的条件信息比较充足，因此三维重建的难度较小，而单一视图的三维重建则比较困难。

三维重建通常采用 4 种表示方式：深度图、体积元素、点云和网格。

1) 深度图

深度图用于表示场景中各点与计算机间的距离，深度图中的每像素表示的是图像中对应的场景与摄像机之间的距离。

2) 体积元素

体积元素又称体素，与像素一样，体素是三维空间内分割的最小的单位，用恒定的标量或向量来表示一个立体的区域。

3) 点云

点云是通过测量仪器得到的图像中物体表面的数据集合。点云可以分为系数点云和密集点云,使用三维坐标测量机得到的间距较大的点云称为稀疏点云,使用三维激光扫描仪得到的比较密集的点云称为密集点云。

如图 3-13(a)、图 3-13(b)、图 3-13(c)分别是前视图、俯视图和左视图,图 3-13(d)是该场景的原图像。从图像中可以看出,该场景包含了简单目标——图片正中间的白色车辆,中等和较难目标——白色汽车左侧的其他车辆。

(a) 前视图　(b) 俯视图　(c) 左视图　(d) 原图像

图 3-13　点云三视图

4) 网格

网格即为用网格模拟组成三维立体物体的表面,计算机视觉中的网格常用的有三角网格和四角网格。

三维重建在实际应用中有不同的方向,例如,自动驾驶和机器人领域中三维重建叫作即时定位与地图构建(Simultaneous Localization And Mapping,SLAM),计算机视觉里还有基于深度学习的三维重建以及对人体的三维重建,对人脸的三维重建,对各种物体的三维重建,对室内场景的三维重建等。

3.4.3　自动驾驶

自动驾驶汽车,也称为无人驾驶汽车,是通过计算机控制来实现的新型技术。自动驾驶是由人工智能、计算机视觉、雷达系统、全球定位系统等多种技术相结合的技术,无须人类的操控即可实现对车辆的安全驾驶。

自动驾驶技术是一项十分复杂、难度极大的工程,因为机动车的驾驶本身就是一件精密度较高的任务,需要驾驶人时刻保持清醒,清晰地观察车辆的周边情况。由于参与交通的因素十分复杂,驾驶人不仅需要观察红绿灯以及四周的车辆,还需要考虑到路上的行人、自行车、电动车、前方道路是否有障碍物、甚至是突然闯入车流的动物,路况信息实时发生改变,稍有不慎就会发生交通事故。

自动驾驶汽车早在 2012 年就已经受到广泛的关注,谷歌自动驾驶汽车于当年的 5 月获得了美国首个自动驾驶车辆的许可证。由于国外地广人稀的明显优势,自动驾驶技术相较于国内发展更为顺利。百度与宝马的自动驾驶研究项目于 2014 年正式开启,并迅速推出了原型车。

2020 年底,北京经济技术开发区建成网联云控式高级别自动驾驶示范区,示范区支持 L4 级别以上的高级别自动驾驶,并且能兼容低级别的自动驾驶。2021 年 5 月举行的第八届国际智能网联汽车技术年会上,北京高级别自动驾驶示范区颁发了国内首批无人配送车车辆编码,并且授予相对应路段的路权,这是我国自动驾驶领域的一次创新突破。图 3-14

和图 3-15 分别为第八届国际智能网联汽车技术年会和无人配送车辆。

图 3-14　第八届国际智能网联汽车技术年会

图 3-15　无人配送车

2021 年 6 月 3 日，广州市为了加强新冠肺炎疫情防控，采取了多种防控措施，多个区域进行全面封闭式管理，该地区的人员只进不出。在广州市委市政府的统一部署下，无人驾驶工程团队连夜对管控区域进行测试，完成无人车的部署。6 月 4 日上午，无人驾驶小巴和无人驾驶出租车驶入疫情管控地区进行物资配送，为区域内的居民提供生活物资。无人驾驶车辆为抗疫工作做出极大贡献，这些车型均不需要配备任何人员，实现了封闭区域内的全无人驾驶，减少了防疫人员的工作量，避免交叉感染的风险，提高了防疫安全性。图 3-16 为防疫期间工作的无人驾驶小巴。

自动驾驶所涉及的技术多种多样，其中十分重要的部分就是计算机视觉，由于车辆驾驶中需要时刻用眼睛观察一切参与交通的要素，因此计算机视觉发挥了它极大的作用，计算机的高算力和低人工成本为自动驾驶提供了坚实的基础。

传统的目标检测特征提取方法对交通场景下不同的目标，包括车辆、行人、路面等都达到了很好的识别效果。自动驾驶涉及多方面的物体识别，其中最为基础的是车辆以及道路的识别，传统的特征提取方法对车道线、马路边缘界限的灰度值以及纹理特征进行处理计

图 3-16　疫情期间的无人驾驶小巴

算,分割出马路的各个区域,但局限性较大。由于马路的视频及图像常受光线、障碍物、树木的阴影、路边杂乱的车辆和行人等的影响,因此传统的简单特征检测方法难以实现复杂路况中的识别任务。自动驾驶的计算机视觉技术经过了长时间的更新迭代,从传统的特征提取方法转为采用深度学习的计算机视觉方法。

深度学习的兴起使得目标识别检测任务的完成质量有了极大的飞跃,在许多情况下甚至在准确度和速度方面超越人类。深度学习的目标检测与传统的检测相比,不仅仅是根据图像中目标表面的特征定位来进行判断,而是进行深入的自主学习。

基于深度学习的自动驾驶通过直接对正确驾驶过程进行学习,来感知实际行驶道路的驾驶方法,对驾驶道路上的路况和目标做整体的判断,而不是局部地对路面、车辆、行人等分别计算,能够极大地提高反应速度。自动驾驶中的计算机视觉任务也包括许多种,例如,车辆定位、三维视觉重建、物体检测分类、语义分割、实例分割、全景分割、运动估计、情景推理、不确定性推理等。因此自动驾驶是一项复杂的复合型工程,为了保证绝对安全的驾驶还需要经过更加周密和严格的测试。

本章小结

随着人工智能与计算机视觉技术的惊人成就在越来越多行业内出现,计算机视觉的未来充满了巨大的希望和想象空间。本章首先介绍计算机视觉的定义和发展,并简单介绍计算机视觉技术的相关学科。由于深度学习与计算机视觉的紧密关系,本章还简单介绍深度学习算法及其对计算机视觉的推动作用。最后,对计算机视觉的关键技术和应用场景进行总结。

本章习题

1. 简述计算机视觉的定义。
2. 数字图像处理和计算机视觉的异同点分别是什么?
3. 现阶段计算机视觉有哪些发展?

4. 举例阐述深度学习的典型算法。
5. 深度学习在计算机视觉中有哪些方面的应用？试举例说明。
6. SIFT 算法的主要步骤有哪些？
7. 调研 Fast R-CNN 和 Faster R-CNN 两种深度学习算法的主要区别。
8. 立体视觉研究的核心是什么？
9. 举例说明计算机视觉在生活中的应用。
10. 人脸识别的主要步骤有哪些？