第 5 章

感 知 智 能

感知智能是指通过各种感觉器官，诸如视觉、听觉、触觉等，与环境进行交互的感知能力。视觉系统使生物体具有视觉感知能力。听觉系统使生物体具有听觉感知能力。利用大数据、深度学习的研究成果，机器在感知智能方面已越来越接近人类水平。

5.1 概述

感知是客观外界直接作用于人的感觉器官而产生的。在社会实践中，人们通过眼、耳、鼻、舌、身 5 个器官能接触客观事物的现象。在外界现象的刺激下，人的感觉器官产生了信息流，沿着特定的神经通道传送到大脑，形成了对客观事物的颜色、形状、声音、冷热、气味、疼痛等的感觉和印象。

感性认识是客观外界直接作用于人的感觉器官而产生的。感性认识在发展中经历感觉、知觉、表象 3 种基本形式。感觉是客观事物的个别属性、特性在人脑中的反映。知觉是各种感觉的综合，是客观事物整体在人脑中的反映，它比感觉全面和复杂。知觉具备选择性、意义性、恒常性以及整体性等特点。在知觉的基础上，产生表象。表象即印象，是通过回忆、联想使这些印象再现出来。它与感觉、知觉不同，是在过去对同一事物或同类事物多次感知的基础上形成的，具有一定的间接性和概括性。但表象只是概括感性材料的最简单的形式，它还不能揭露事物的本质和规律。

视觉在人类的感觉世界中担负着重要的任务。我们对大部分环境信息作出反应，是经过视觉传入大脑的。它在人类的感觉系统中占主导地位。如果人类用视觉接收一个信息，而另外一个信息是通过另一个感觉器官接收的，又如果这两个信息相互矛盾，人们所反应的一定是视觉信息。

20 世纪 80 年代，按照马尔的视觉计算理论，计算机视觉分 3 个层次处理。

(1) 对图像进行边缘检测与图像分割等底层视觉处理。

(2) 求取深度信息、表面朝向等二维半描述，主要方法有：由影调、轮廓、纹理等恢复三维形态；由体视恢复景物的深度信息；由图像序列分析确定物体的三维形状和运动参数；距离图像获取与分析；结构光方法等。

(3) 根据三维信息对物体进行建模、表示与识别，可采用基于广义圆柱体的方法。另一常用方法是将物体外形表示为平面或曲面块(简称面基元)的集合，每个面基元的参数以及面基元之间的相互关系用属性关系结构来表示，从而将物体识别问题转化为属性关系结构的匹配问题。

1990年,阿罗莫讷斯(J. Aloimonos)提出定性视觉、主动视觉等。定性视觉方法的核心是将视觉系统看成执行某一任务的更大系统的子系统,视觉系统所要获取的信息,只是完成大系统任务所必需的信息。主动视觉方法则集感知、规划与控制为一体,通过这些模块的动态调用和信息获取过程与处理过程的相互作用,来更有效地完成视觉任务。该方法的核心是主动感知机制的建立,就是根据当前任务、环境状况、阶段处理结果和有关知识,来规划和控制下一步获取信息的传感器类型及其位姿。实现多视点或多传感器的数据融合,也是其关键技术。

听觉过程包括机械→电→化学→神经冲动→中枢信息加工等环节。20世纪80年代,有关语音识别和语言理解的研究得到了很大的加强和发展。美国国防部高级项目管理局自1983年开始为期10年的DARPA战略计算工程项目,其中包括用于军事领域的语音识别和语言理解、通用语料库等。

IBM使用离散参数HMM(隐马尔可夫模型)构成一些基本声学模型,然后利用固定的有限个基本声学模型构成字(word)模型。这种方法可以利用较少的训练数据获得较好的统计结果。并且,这种方法可以使训练自动完成。

进入20世纪90年代,神经网络成为语音识别的一条新途径。人工神经网络(ANN)具有自适应性、并行性、非线性、鲁棒性、容错性和学习特性,在结构和算法上都显示出其实力,它可以联想模式对,将复杂的声学信号映射为不同级别的语音学和音韵学的表示,不必拘束于选取特殊的语音参数,而对综合的输入模式进行训练和识别,可把听觉模型融于网络模型之中。

2006年,欣顿(G. E. Hinton)等提出深度学习。2010年,欣顿使用深度学习搭配GPU的计算,使语音识别的计算速度提升了70倍以上。2012年,深度学习出现新一波高潮,那年的ImageNet大赛(有120万张照片作为训练组,5万张作为测试组,要进行1000个类别分组)首次采用深度学习,把过去好几年只有微幅变动的错误率,一下由26%降低到15%。而同年微软团队发布的论文中显示,他们通过深度学习将ImageNet 2012数据集的错误率降到了4.94%,比人类的错误率5.1%还低。2015年,微软再度拿下ImageNet 2015冠军,此时错误率已经降到了3.57%的超低水平。微软用的是152层深度学习网络。

基于视觉、听觉等感知能力的感知智能近年来取得了重要进展,在业界多项权威测试中,人工智能系统都已经达到甚至超过人类水平,感知智能正迎来它最好的时代。人脸识别、语音识别等感知智能技术如今已运用在图片处理、安防、教育、医疗等多个领域。

5.2 知觉理论

知觉理论是指人类系统地对环境信息加以选择和抽象概括的理论。迄今为止,主要建立了4种知觉理论:建构理论、格式塔理论、直接知觉理论、拓扑视觉理论。

5.2.1 建构理论

过去的知识经验主要是以假设、期望或因式的形式在知觉中起作用。人在知觉时接收感觉输入,在已有经验的基础上,形成关于当前的刺激是什么,或者激活一定的知识单元而形成对某种客体的期望。知觉是在这些假设、期望等的引导和规划下进行的。布鲁纳(J. S. Bruner)等发展建构理论,认为所有感知都受到人们的经验和期望的影响。建构理论的基本假设如下。

(1) 知觉是一个活动的、建构的过程,它在某种程度上要多于感觉的直接登记,……其他事件会切入刺激和经验中来。

(2) 知觉并不是由刺激输入直接引起的,而是所呈现刺激与内部假设、期望、知识以及动机和情绪因素交互作用的终极产品。

(3) 知觉有时会受到不正确的假设和期望的影响,因而知觉也会发生错误。

建构理论关于知觉的看法是把记忆的作用赋予极大的重要性。他们认为先前经验的记忆痕迹,加到此时此地被刺激诱导出来的感觉中,因此就构造出一个知觉象。而且,建构论者主张有组织的知觉基础是从一个人的记忆中选择、分析并添加刺激信息的过程,而不是格式塔论者所主张的大脑组织的天生定律所引起的自然操作作用。

知觉的假设考验说是一种建立在过去经验作用基础上的知觉理论。支持这个理论的还有其他的重要论据。例如,外部刺激与知觉经验并没有一对一的关系,同一刺激可引起不同的知觉,不同的刺激却又可以引起相同的知觉。知觉是定向、抽取特征,与记忆中的知识相对照,然后再定向、再抽取特征并再对照,如此循环,直到确定刺激的意义,这与假设考验说有许多相似之处。

5.2.2 格式塔理论

格式塔(Gestalt)心理学诞生于 1912 年。格式塔心理学家发现的感知组织现象是一种非常有力的关于像素整体性的附加约束,从而为视觉推理提供了基础。格式塔是德文 Gestalt 的译音。英文中常译成 form(形式)或 shape(形状)。格式塔心理学家所研究的出发点是"形",它是指从由知觉活动组织成的经验中的整体。换言之,格式塔心理学家认为任何"形"都是知觉进行了积极组织或构造的结果或功能,而不是客体本身就有的。它强调经验和行为的整体性,反对当时流行的建构主义元素学说和行为主义"刺激-反应"公式,认为整体不等于部分之和,意识不等于感觉元素的集合,行为不等于反射弧的循环。尽管格式塔原理不只是一种知觉的学说,但它却来源于对知觉的研究,而且一些重要的格式塔原理,大多是由知觉研究所提供的。格式塔派学者们相信大脑中组织之固有和天生的法则。他们辩论说,这些法则就解释了这些重要现象:图形——背景的分化、对比、轮廓线、趋合、知觉组合的原则以及其他组织上的事实。格式塔派学者们认为,在他们所提出的各种知觉因素之后存在着一个"简单性"原则。他们断言,包含着较大的对称性、趋合、紧密交织在一起的单位以及相似的单位的任何模式,对于观察者来说,外表上显得"比较简单"。如果一个构造可以有一种以上的方式看到,例如,一个线条构成的图画可以看成是扁平的或者一个正方块,那个"较简单的"方式会更通常一些。格式塔派学者们并没有忽视潜在经验对于知觉的效应,但是他们的首要着重点是放在成为神经系统不可分的内在机制的作用上。因此,他们假设,似动或 Φ 现象是大脑天生组织起来倾向的结果。

单个图形背景的模式一般很少,典型的模式是几个图形有一个共同的背景。一些单个的图形还倾向于被知觉集聚在一起的不同组合。格式塔心理学创始人之一的韦特海姆系统地阐述了如下"组合原则"。

(1) 邻近原则。彼此紧密邻近的刺激物比相隔较远的刺激物有较大的组合倾向。邻近可能是空间的,也可能是时间的。按不规则的时间间隔发生的一系列轻拍响声中,在时间上接近的响声倾向于组合在一起。邻近而组合成的刺激不必都是同一种感觉形式的。例如,夏天下雨时,雷电交加,我们就把它们知觉为一个整体,即知觉为同一事件的组成部分。

(2) 相似原则。彼此相似的刺激物比不相似的刺激物有较大的组合倾向。相似意味着强度、颜色、大小、形状等这样一些物理属性上的类似。俗话说,"物以类聚,人以群分",也就包含

这种原则。

(3) 连续原则。人们知觉倾向于知觉连贯或连续流动的形式,即一些成分和其他成分连接在一起,以便有可能使一条直线、一条曲线或者一个动作沿着已经确立的方向继续下去。

(4) 闭合原则。人们知觉倾向于形成一个闭合或更加完整的图形。

(5) 对称原则。人们知觉倾向于把物体知觉为一个中心两边的对称图,导致对称或平衡的整体而不是非对称的整体。

(6) 共方向原则。也称共同命运原则。如果一个对象中的一部分都向共同的方向去运动,那这些共同移动的部分就易被感知为一个整体。这个组合原则本质上是相似组合在运动物体上的应用,它是舞蹈设计中的一个重要手段。

在每一种刺激模式中,一些成分都有某种程度的接近、某种程度的类似以及某种程度适合"好图形"的东西。有时组合的一些倾向在同一方向上起作用,有时它们彼此冲突。例如,图 5.1 给出了格式塔知觉组织原则例图。

图 5.1 格式塔知觉组织原则例图

格式塔心理学家试图根据心脑同形观来解释知觉原则。按照这种心脑同形观,视觉组织经验与大脑中的某一过程严格对应。当我们观察环境时,格式塔心理学家假定大脑中存在一种电场,以帮助产生相对稳定的知觉组织经验。格式塔心理学家主要依赖内省报告或"注视一个图形并从你自己的角度观看"的方法研究知觉。不幸的是,格式塔心理学家对大脑的工作机制知之甚少,而且他们的虚拟生物学解释也没有得到承认。

格式塔理论反映了人类视觉本质的某些方面,但它对感知组织的基本原理只是一种公理性的描述,而不是一种机理性的描述。因此,自从 20 世纪 20 年代提出以来未能对视觉研究产生根本性的指导作用,但是研究者对感知组织原理的研究一直没有停止。特别是在 20 世纪 80 年代以后,威特肯(Witkin)、坦丁鲍姆(Tenenbaum)、劳卫(Lowe)和蓬特兰德(Pentland)等在感知组织的原理,以及在视觉处理中的应用等方面取得了新的重要研究成果。

5.2.3 直接知觉理论

美国心理学家吉布森(J. J. Gibson)因其对知觉的研究而闻名于学术界。1950 年他提出生态知觉理论,认为知觉是直接的,没有任何推理步骤、中介变量或联想。生态学理论(刺激物说)与建构理论(假设考验说)相反,主张知觉只具有直接性质,否认已有知识经验的作用。吉布森认为,自然界的刺激是完整的,可以提供非常丰富的信息,人完全可以利用这些信息,直接产生与作用于感官的刺激相对应的知觉经验,根本不需要在过去经验基础上形成假设并进行考验。根据他的生态知觉理论,知觉是和外部世界保持接触的过程,是刺激的直接作用。他把这种直接的刺激作用解释为感官对之作出反应的物理能量的类型和变量。知觉是环境直接作用的产物这一观点,是和传统的知觉理论相背离的。

吉布森的知觉理论之所以被冠之以"生态知觉理论",原因在于它强调与生物适应最有关

系的环境事实。对吉布森而言，感觉是因演进而对环境的适应，而且环境中有些重要现象，如重力、昼夜循环和天地对比等，在进化史上都是不变的。不变的环境带来稳定性，并且提供了个体生活的参照框架。因此，种系演化的成功依靠正确地反映环境的感觉系统。从生态学的观点来看，知觉是环境向知觉者显露的过程，神经系统并非建构知觉，而是萃取它们。

吉布森认为知觉系统从流动的系列中抽取不变性。他的理论现在称作知觉的生态学理论，并形成了一个学派，主要假设如下。

(1) 刺激眼睛的光线模式是一个光学分布(optic array)。这种结构性的光线包含来自环境中的所有投射到眼睛的视觉信息。

(2) 这种光学分布提供关于空间中目标分布特征的明确的或恒定的信息。这种信息存在多种形式，包括结构极差、光流模式和功能承受性。

(3) 知觉是在很少或没有信息加工参与的情况下，通过共振直接从光学分布中提取各种丰富信息。

吉布森把具有结构的表面的知觉称为正常的或生态学的知觉。他认为，与他自己的看法相比，格式塔理论主要以特殊情况下的知觉分析为根据，在这种情况下，结构化减少了或者是毫不相干的，就像这张纸的结构对于印在上面的内容毫不相干一样。

在构造论理论中，知觉常常是被用来自记忆的信息。而吉布森认为具有结构表示的高度结构起来的世界提供了足够丰富而精确的信息，观察者可以从中选择，而无须再从过去储存起来的信息中选择。生态学理论坚信人们都是用相似的方法去看待世界，高度重视在自然环境中可得到的信息的全面复合的重要性。

吉布森的生态知觉理论具有一定的科学依据。他假设知觉反应是天生的观点与新生动物的深度知觉是一致的，同时也符合神经心理学中视觉皮层单一细胞对特定视觉刺激有所反应的研究结论。但是，他的理论过分强调个体知觉反应的生物性，忽视了个体经验、知识和人格特点等因素在知觉反应中的作用，因而也受到了一些研究者的批评。

建构理论与吉布森范式的区别之一是前者重视自上而下加工在知觉中的作用，而后者则强调自下而上加工的重要性。事实上，自上而下加工和自下而上加工对知觉的相对重要性取决于不同因素的影响。当观察条件良好时，视知觉主要由自下而上加工决定，但是当快速呈现刺激或刺激清晰度不够导致观察条件不理想时，视知觉主要涉及自上而下加工过程。与以上分析一致的是，吉布森重点考察优化条件下的视知觉，而建构主义则常常选用一些不太理想的观察条件进行知觉研究。

间接和直接理论存在很大的区别，因为相关的理论家所追求的目标很不相同。如果我们考虑针对识别的知觉和针对行动的知觉之间的区别，这一点就会明朗得多。来自认知神经科学和认知神经心理学的证据也支持二者之间存在区别这一观点。这方面的证据表明一条腹侧加工通路更多地参与针对识别的知觉，而一条背侧加工通路更多地参与针对行动的知觉。绝大多数知觉理论家都集中在探讨针对识别的知觉上，而吉布森则强调针对行动的知觉。

5.2.4　拓扑视觉理论

在视知觉研究 200 多年的历史中，始终贯穿着“原子论”和“整体论”之争。原子论认为，知觉过程开始于对物体的特征性质或简单组成部分的分析，是从局部性质到大范围性质。而整体论却认为，知觉过程开始于物体的整体性的知觉，是从大范围性质到局部性质。

1982年,陈霖在《科学》杂志上就知觉过程从哪里开始的根本问题,原创性地提出了“拓扑性质初期知觉”的假说。这是他在视知觉研究领域的独创性贡献,向半个世纪以来占统治地位的初期特征分析理论提出了挑战。与传统的初期特征分析理论根本不同,拓扑性质初期知觉理论从大范围性质到局部性质的不变性知觉的角度,为理解知觉信息基本表达的问题,为理解知觉和认知过程的局部和整体的关系问题,为理解认知科学的理论基础——认知和计算的关系问题,提出了一个理论框架。

一系列视知觉实验表明,视图形知觉有一个功能层次,视觉系统不仅能检测大范围的拓扑性质,而且较之局部几何性质视觉系统更敏感于大范围的拓扑性质,对由空间相邻关系决定的大范围拓扑性质的检测是发生在视觉时间过程的最初阶段。

拓扑学研究的是在拓扑变换下图形保持不变的性质和关系,这种性质和关系就称为拓扑性质。所谓拓扑变换是一对一的连接变换,它可以形象地想象成橡皮薄膜的任意变形,只要不把薄膜剪开或不把薄膜的任意两点粘合起来。一张橡皮薄膜可以任意地变形,可以从一个三角形变成一个正方形,三角形可以变成圆形或任意不规则的图形(见图5.2),只要不把它剪开。作为一个连通的整体这个性质,即连通性,仍然保持不变的。所以,连通性是一种拓扑性质。另外,一个连通的图形中有没有洞或者有几个洞,这种性质也是一种典型的拓扑性质。

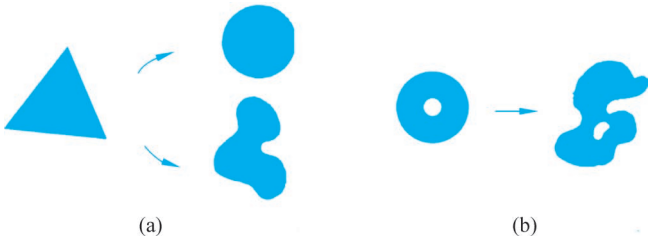

图5.2 拓扑变换和拓扑性质的图示

根据人们的直觉的经验,圆、三角形和正方形看起来是很不相同的图形,但是从拓扑学的角度来看,它们都是拓扑等价的、相同的。而圆和环,由于一个含有一个洞,另一个不含有洞,它们是拓扑不同的。尽管在通常的视觉观察的条件下,从人们在心理学上相似性的角度来说,人们会觉得圆和环比较比圆和三角形、正方形要相像一些,但是如果视觉系统具有初期提取拓扑性质的功能,那么我们应当预计,在不能把圆和三角形、正方形区别开来的短暂呈现的条件下,却仍然有可能把圆和环区别开来。图5.3表示用于这类实验的三组刺激图形,它们分别是实心圆和实心正方形、实心圆和实心三角形、实心圆和环。受实验者被要求注视每幅图的中心的黑点,然后每幅图被呈现短暂的5ms,并且在撤去之后立即呈现另一幅空白的没有图形的蔽掩刺激,来干扰视觉系统对在此以前呈现的图形的知觉。受实验者被要求回答的问题并不是被呈现的在注视点两旁的图形是什么样的图形,而是被呈现的两个图形是一样的或是不一样的。

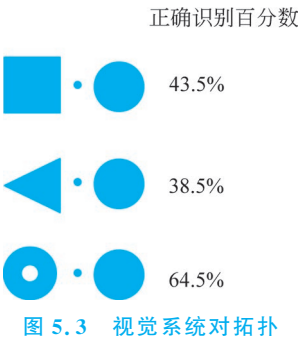

图5.3 视觉系统对拓扑差异的敏感性

实验的结果也表示在图5.3中。主要的实验发现是,视觉系统确实更敏感于拓扑性质的差异,也就是敏感于具有一个洞的环和没有洞的实心圆的差别。对圆和环一组刺激图形的正确报告率,要显著高于圆和三角形的正确报告率与圆和正方形的正确报告率。而且,拓扑性质等价的两对图形,圆和三角形与圆和正方形,它们的正确报告率的区别却没有达到统计意义,从而作为对照实验加强了视觉系统对圆和环的差别的敏感就是对它们之间

的拓扑差异敏感的假设。这个同日常经验不一致却跟拓扑学的解释一致的实验，提供了一个支持拓扑结构假设的较为直接和令人信服的证据。

2005年，陈霖在 *Visual Cognition* 第四期上发表长达88页的“重大主题论文”，对拓扑视觉理论概括为：知觉组织的拓扑学研究基于一个核心思想并包括两方面。核心思想是，知觉组织应该从变换(transformation)和变换中的不变性(invariance)知觉的角度来理解。第一方面，强调形状知觉中的拓扑结构，这就是，知觉组织的大范围性质能够用拓扑不变性来描述；第二方面，进一步强调早期拓扑性质知觉，这就是拓扑性质知觉优先于局部特征性质的知觉。“优先”有两个严格的含义：一是由拓扑性质决定的整体组织是知觉局部几何性质的基础；二是基于物理连通性的拓扑性质知觉先于局部几何性质的知觉。

5.3 视觉感知

5.3.1 视觉通路

视觉系统使生物体具有视觉感知能力。它使用可见光信息构筑机体对周围世界的感知。根据图像发现周围景物中有什么物体和物体在什么地方的过程，也就是从图像得到对观察者有用的符号描述的过程。视觉系统具有将外部世界的二维投射重构为三维世界的能力。需要注意的是，不同物体所能感知的可见光处于光谱中的不同位置。

光线进入眼到达视网膜。视网膜是脑的一部分，由处理视觉信息的几种类型的神经元组成的。它紧贴在眼球的后壁上，厚度只有0.5mm左右，包括三级神经元：第一级是光感受器，由无数视杆细胞和视锥细胞组成；第二级是双极细胞；第三级是神经节细胞。由神经节细胞发出的轴突形成视神经。这三级神经元构成了视网膜内视觉信息传递的直接通道。

视网膜内有4种光感受器：视杆细胞和3种视锥细胞。在每一种感受器内都含有一种特殊的色素。当一个这样的色素分子吸收了一个光量子以后，它会在细胞内触发一系列的化学变化；与此同时释放出能量，导致电信号的产生和突触化学递质的分泌。视杆细胞的视色素称“视紫红质”，其光谱吸收曲线的峰值波长为500nm。3种视锥细胞色素的光谱吸收峰值分别为430nm、530nm和560nm，分别对蓝、绿、红3种颜色最敏感。

视神经在进入脑中枢前以一种特殊的方式形成交叉，即从两眼鼻侧视网膜发出的纤维交叉到对侧大脑半球；从颞侧视网膜发出的纤维不交叉，投射到同侧大脑半球。其结果是：从左眼颞侧视网膜来的纤维和从右眼鼻侧来的纤维汇聚成左侧视束，投射到左侧外膝体；再由左外膝体投射到左侧大脑半球，与相应脑区对应的是右侧半个视野。相反地，从左眼鼻侧视网膜来的纤维和从右眼颞侧视网膜来的纤维汇聚成右侧视束，投射到右侧外膝体；再由右侧外膝体投射到右侧半球，相应脑区对应于左侧半个视野。脑两个半球的视皮层通过胼胝体的纤维互相连接。这种相互连接，使从视野两边得来的信息混合起来。

视皮层本身的神经元主要有两种：星形细胞和锥体细胞。星形细胞的轴突与投射纤维形成联系。锥体细胞呈三角形，尖端朝表层，向上发出一个长的树突，基底则发出几个树突作横向联系。

视皮层和其他皮层区一样，包括6个细胞层次，由表及里用罗马数字Ⅰ～Ⅵ来代表。皮层神经元的突起(树突和轴突)的主干都沿与皮层表面相垂直的方向分布；树突和轴突的分枝则横向分布在不同层次内。不同皮层区之间由轴突通过深部的白质进行联系，同一皮层区内由树突或轴突在皮层内的横向分枝来联系。

近年来，视皮层的范围已扩大到顶叶、颞叶和部分额叶在内的许多新皮层区，总数达25个。另外，还有7个视觉联合区，这些皮层区兼有视觉和其他感觉或运动功能。所有视区加在一起占大脑新皮层总面积的55%。由此可见视觉信息处理在整个脑功能中所占有的分量。研究各视区的功能分工、等级关系以及它们之间的相互作用，是当前视觉研究的一个前沿课题。确定一个独立的视皮层区的依据是：①有独立的视野投射图，该区与其他皮层区之间有相同的输入和输出神经联系；②该区域内有相似的细胞筑构；③有不同于其他视区的功能特性。

韦尼克(Wernicke)和格什温德(Geschwind)认为，视觉识别的神经通路如图5.4所示。根据他们的模型，视觉信息由视网膜传至外侧膝状体，从外侧膝状体传至初级视皮层(17区)，然后传至一个更高级的视觉中枢(18区)，并由此传至角回，然后至Wernicke区。在Wernicke区，视觉信息转换为该词的语声(听觉)表象。声音模式形成后，经弓状束传至Broca区。

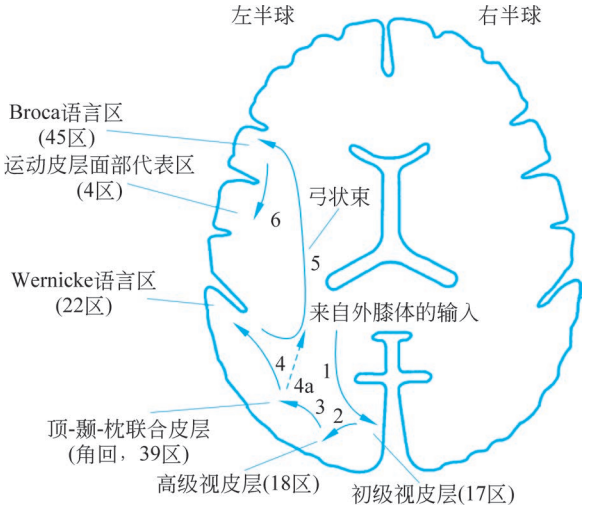

图5.4 视觉的神经通路

视皮层中17区被称为第一视区(V1)或纹状皮层。它接受外侧膝状体的直接输入，因此也称为初级视皮层。对视皮层的功能研究大多数是在这一级皮层进行的。除了接受外侧膝状体直接投射的17区之外，和视觉有关的皮层还有纹前区(18区)和纹外区(19区)。根据形态和生理学的研究，17区不投射到侧皮层而仅射到18区，18区向前投射到19区，但又反馈到17区。18区内包括3个视区，分别称为V2、V3和V3A，它们的主要输入来自V1。V1和V2是面积最大的视区。19区深埋在上颞沟后壁，包括第四(V4)和第五视区(V5)。V5也称作中颞区，已进入颞叶范围。颞叶内其他与视觉有关的皮层区还有内上额区、下颞区。顶叶内有顶枕区、腹内顶区、腹后区和7a区。枕叶以外的皮层区可能属于更高的层次。为什么要这样多的代表区？是不是不同代表区检测图形的不同特征(如颜色、形状、亮度、运动、深度等)？或是不同代表区代表处理信息的不同等级？会不会有较高级的代表区把图形的分离特征整合起来，从而给出图形的生物学含义？是不是有专门的代表区负责储存图像(视觉学习记忆)或主管视觉注意？这些都将是在一个更长的时间内视觉研究有待解决的问题。

视皮层神经元对光点刺激的反应很弱，只有在感受野内用适当方位(朝向)的光点给予刺激才能引起兴奋。根据皮层神经元感受野结构的不同，休贝尔(Hubel)和维塞勒(Wiesel)对猫和猴的视皮质中单一神经元的激发模式进行了研究，发现有4种类型视皮层神经元——简

单细胞、复杂细胞、超复杂细胞和极高度复杂细胞。

知觉恒常性是指人能在一定范围内不随知觉条件的改变而保持对客观事物相对稳定特性的组织加工的过程。它是人们知觉客观事物的一个重要的特性。

视觉感知主要有两个功能：一是目标知觉(即它是什么)；二是空间知觉(即它在哪里)。已有确实的证据表明，不同的大脑系统分别参与上述两种功能。如图 5.5 所示，腹部流从视网膜开始，沿腹部经过侧膝体(LGN)、初级视网皮层区域(V1、V2、V4)、下颞叶皮层(IT)，最终到达腹外侧额叶前部皮层(VLPFC)，主要处理物体的外形轮廓等信息，即主要负责物体识别；背部流从视网膜开始，沿背部流经过侧膝体(LGN)、初级视皮层区域(V1、V2)、中颞叶区(MT)、后顶叶皮层(PP)，最后到达背外侧额叶前部皮层(DLPFC)，主要处理物体的空间位置信息等，即处理负责物体的空间定位等。因此，这两条信息流也被称为 what 通路和 where 通路。

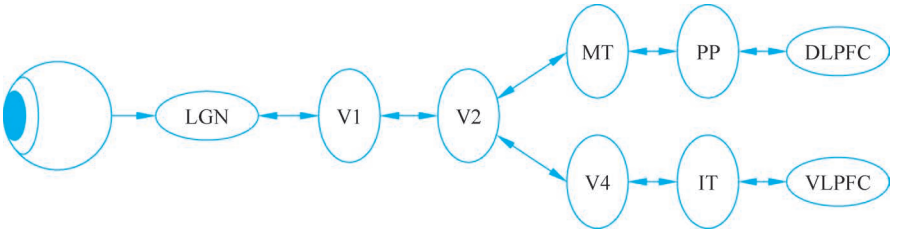

图 5.5　视觉感知通路

5.3.2　马尔的视觉计算理论

马尔(D. Marr)在 20 世纪 70 年代末 80 年代初创立了视觉的计算理论，使视觉的研究前进了一大步。马尔的视觉计算理论立足于计算机科学，系统地概括了心理物理学、神经生理学、临床神经病理学等方面已取得的所有重要成果，是迄今为止最系统的视觉理论。马尔理论的出现对神经科学的发展和人工智能的研究产生了深远的影响。

视觉是一个信息处理过程。这个过程根据外部世界的图像产生对观察者有用的描述。这些描述依次由许多不同但固定的、每个都记录了外界的某方面特征的表象(representation)所构成或组合而成。一种新的表象之所以提高了描述是因为新的表象表达了某种信息，而这种信息将便于对信息作进一步解释。按这种逻辑来思考可得到这样的结论，即在对数据作进一步解释以前我们需要关于被观察物体的某些信息，这就是所谓的本征图像。然而，数据进入我们的眼睛是要以光线为媒介的。灰度图像中至少包含关于照明情况、观察者相对于物体位置的信息。因此，按马尔的方法首先要解决的问题是如何把这些因素分解开。他认为低层视觉(即视觉处理的第一阶段)的目的就是要分清哪些变化是由哪些因素引起的。大体上来说，这个过程要经过两个步骤来完成：第一步是获得表示图像中变化和结构的表象。这包括检测灰度的变化、表示和分析局部的几何结构，以及检测照明的效应等处理。第一步得到的结果被称为初始简图(Primal Sketch)的表象；第二步对初始简图进行一系列运算得到能反映可见表面几何特征的表象，这种表象被称为二维半(2.5 D)简图或本征图像。这些运算中包括由立体视觉运算提取深度信息，根据灰度影调、纹理等信息恢复表面方向，由运动视觉运算获取表面形状和空间关系信息等。这些运算的结果都集成到本征图像这个中间表象层次。因为这个中间表象已经从原始的图像中去除了许多的多义性，是纯粹地表示了物体表面的特征，其中包括光照、反射率、方向、距离等。根据本征图像表示的这些信息可以可靠地把图像分成有明确含义

的区域(这称为分割),从而可得到比线条、区域、形状等更为高层的描述。这个层次的处理称为中层视觉处理(intermediate processing)。马尔视觉理论中的下一个表象层次是三维模型,它适用于物体的识别。这个层次的处理涉及物体,并且要依靠和应用与领域有关的先验知识来构成对景物的描述,因此被称为高层视觉处理。

马尔的视觉计算理论虽然是首次提出的关于视觉的系统理论,并已对计算机视觉的研究起了巨大的推动作用,但还远未解决人类视觉的理论问题,在实践中也已遇到了严重困难。对此已有不少学者提出改进意见。

马尔首先研究了解决视觉理解问题的策略。他认为视觉是一个信息处理问题,它需要从3个层次来理解和解决。

(1) 计算理论层次——研究对什么信息进行计算和为什么要进行这些计算。

(2) 表示和算法层次——实际执行由计算理论所规定的处理,输入输出如何表示,以及将输入变换到输出的算法。

(3) 硬件实现——实现由表示和算法层次所考虑的表示,实现执行算法,研究完成某一特定算法的具体机构。

例如,傅里叶变换是属于第一层的理论,而计算傅里叶变换的算法(如快速傅里叶变换算法)是属于第二个层次的。至于实现快速傅立叶算法的阵列处理机就属于硬件执行的层次。可以认为,视觉是一个过程,这个过程从外部世界的图像产生对观察者有用的描述。这些描述依次地由许多不同的、但是固定的、每个都记录了景物的某方面的表示法所构成或组合而成。因此选择表示法对视觉的理解是至关重要的。根据马尔所提出的假设,视觉信息处理过程包括3个主要表示层次:初始简图、二维半简图和三维模型。根据某些心理学方面的证据,人类视觉系统的表示法如图5.6所示。

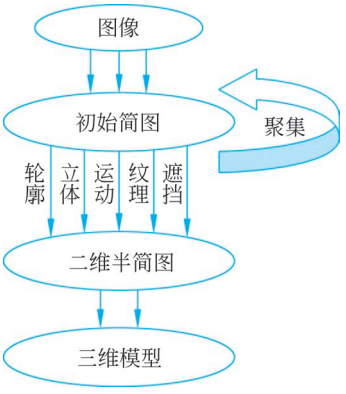

图5.6 视觉系统的表示层次

1. 初始简图

在灰度图像中,包含两种重要的信息:图像中存在的灰度变化和局部的几何特征。初始简图是一种基元表示法,它可以完全而清楚地表示这些信息。初始简图所包含的大部分信息集中在与实际的边缘以及边缘的终止点有关的急剧的灰度变化上。每个由边缘引起的灰度变化,在初始简图上都有相应的描述。这样的描述包括:与边缘有关的灰度变化率,总的灰度变化,边缘的长度、曲率以及方向。粗略地说,初始简图是以勾画草图的形式来表示图像中的灰度变化。

2. 二维半简图

图像中的灰度受多种因素的影响,其中主要包括光照条件、物体几何形状、表面反射率以及观察者的视角等。因此,先要分清上述因素的影响,也就是对景物中物体表面作更充分的描述,才能着手建立物体的三维模型,这就需要在初始简图与三维模型之间建立一个中间表示层次,即二维半简图。物体表面的局部特性可以用所谓的内在特性来描述。典型的内在特性包括表面方向、观察者到表面的距离,反射和入射光照、表面的纹理和材料特性。内在图像由图像中各点的某项单独的内在特性值,以及关于这项内在特性在什么地方产生不连续的信息所组成(见表5.1)。二维半简图可以看成某些内在图像的混合物。简而言之,二维半简图完全而清楚地表示关于物体表面的信息。

表 5.1　二维半简图

信　息　源	信 息 类 型	信　息　源	信 息 类 型
立体视觉	视差，因而可得到 $\delta\gamma$、$\Delta\gamma$ 和 S	其他遮挡线索	$\Delta\gamma$
方向选择性	$\Delta\gamma$	表面方向轮廓	Δs
从运动恢复结构	γ，$\delta\gamma$，$\Delta\gamma$ 和 S	表面纹理	可能有 γ
光源	γ 和 S	表面轮廓	$\Delta\gamma$ 和 S
遮挡轮廓	$\Delta\gamma$	影调	δs 和 Δs

注：γ 相对深度（按垂直投影），就是观察者到表面点的距离；$\delta\gamma$、γ 的连续或小的变化；$\Delta\gamma$，γ 的不连续点；S 局部表面方向；δ_s，S 的连续或小的变化；ΔS，S 的不连续点。

在初始简图和二维半简图中，信息经常是以和观察者联系在一起的坐标为参考表示的，因此这种表示法被称为是以观察者为中心的表示法。

3. 三维模型

在三维模型表象中，以一个形状的标准轴线为基础的分解最容易得到。在这些轴线中，每条轴线都和一个粗略的空间关系相联系，这种关系对包含在该空间关系范围内的主要的形状组元轴线提供了一种自然的组合方式。用这种方法定义的模块称为三维模型。所以，每一个三维模型具有以下模块。

（1）一根模型轴，指的是能确定这一模型的空间关系的范围的单根轴线。它是表象的一个基元，能粗略地告诉我们被描述的整体形状的若干性质，例如，整体形状的大小信息和朝向信息。

（2）在模型轴所确定的空间关系机含有主要组元轴的相对空间位型和大小尺寸可供选择。组元轴的数目不宜太多，它们的大小也应当大致相同。

（3）一旦和组元轴相联系的形状组元的三维模型被构造出来，那么就可以确定这些组元的名称（内部关系）。形状组元的模型轴对应于这个三维模型的组元轴。

在图 5.7 中，每一个方框都表示一个三维模型，模型轴画在方框的左侧，组元轴则画在右侧。人体三维模型的模型轴是一基元，它把整个人体形状的大体性质（大小和朝向）表达清楚。对应于躯干、头部、肢体的 6 根组元轴各自可以和一个三维模型联系起来，这种三维模型包含着进一步把这些组元轴分解成更小的组元构型的附加信息。尽管单个三维模型的结构很简单，但按照这种层次结构把几个模型组合起来，就能在任意精确的程度上构成二种能抓住这一形状的几何本质的描述。我们把这种三维模型的层次结构称为一个形状的三维模型描述。

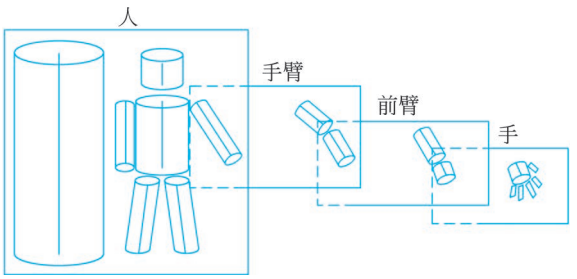

图 5.7　人的三维模型

三维表示法完全而清楚地表示有关物体形状的信息。采用广义柱体的概念虽然很重要，却很简单。一个普通的圆柱可以看成是一个圆沿着通过它的中心线移动而形成的。更一般的情况，一个广义柱体是二维的截面沿着称为轴线移动而成。在移动过程中，截面与轴之间保持固定的角度。截面可以是任何形状，在移动过程中它的尺寸可能是变化的，轴线也不一定是直线。

5.3.3 图像理解

图像理解(image understanding,IU)就是对图像的语义理解,用计算机系统解释图像,实现类似人类视觉系统理解外部世界的对象,理解图像中的目标、关系、场景,能回答该图像的“语义”内容的问题,例如,画面上有没有人?有几个人?每个人在做些什么?图像理解一般可以分为4个层次:数据层、描述层、认知层和应用层。各层的主要功能如下。

(1) 数据层:获取图像数据,这里的图像可以是二值图、灰度图、彩色的和深度图等。主要涉及图像的压缩和传输。数字图像的基本操作如平滑、滤波等一些去噪操作也可归入该层。该层的主要操作对象是像素。

(2) 描述层:提取特征,度量特征之间的相似性(即距离),采用的技术有子空间方法(Subspace),如ISA、ICA、PCA。该层的主要任务就是将像素表示符号化(形式化)。

(3) 认知层:图像理解,即学习和推理(Learning and Inference),该层是图像理解系统的“发动机”。该层非常复杂,涉及面很广,正确的认知(理解)必须有强大的知识库作为支撑。该层操作的主要对象是符号。具体的任务还包括数据库的建立。

(4) 应用层:根据任务需求实现分类、识别、检测,设计相应的分类器、学习算法等。

图像理解的主要研究内容包括目标识别、高层语义分析及场景分类等。

1. 目标识别

让计算机识别判断场景中有什么物体,在哪儿,解决“what-where”问题,这是计算机视觉的主要任务,也是图像理解的基本任务。场景中的“目标”通常可视为具有较高显著度并符合局部感知一致性的区域,目标识别的过程也是计算机对场景中的物体进行特征分析和概念理解的过程。通常地,目标识别的整个过程包括了目标判断、目标分类和目标定位,目标判断分析场景中是否存在指定类别的目标;目标分类分析划定的目标区域是何种类别;目标定位确定目标在场景中的位置,定位中的目标检测基于区域表述,用规则形状(矩形或圆)标记目标区域,而像素级别的目标定位则通过视觉分割从场景中提取完整的目标区域。

2. 高层语义分析

图像理解是通过计算机对输入场景的计算、分析和推理将场景的相应目标和区域进行语义化标记输出的过程,因此高层语义分析对图像理解的实现具有重要作用。由于对目标和场景进行了认知上的概念划分,因此只要有足够的训练学习均可将其进行简单的名称语义化描述。更通常的语义化描述则涉及通用的概念模型描述,并建立区域特征与语义单词的概率对应关系,体现了数据和知识概念转换,研究侧重于视觉的中低层数据特征的分析提取和概率关系建模,一定程度上实现了自动的语义标记。

由于样本获取和概念描述的多义性等影响,图像语义化研究仅仅处于初始阶段,主要以检索语义化为主,各种语义化的标记过程对概念区域的描述非常有限,数据和知识的对应关系通常设计模型进行参数化学习和概率分析,最大后验概率得到的对应关系就是最终语义化的结果。也可通过建立知识模型对匹配推理得到的结果进行语义化标记。

3. 场景分类

场景分类是图像理解中对整体场景的判断和解释。2006年在MIT首次召开了场景理解研讨会(scene understanding symposium,SUNS),明确了场景分类将会是图像理解一个新的有前途的研究热点。目前,对场景分析的研究集中于视觉心理学和生理学,快速场景感知试验证明人无须感知场景中的目标便可通过空间布局分析语义场景内容,对场景理解仅需很短的

时间便获取到大量的信息，从眼睛获取到的视觉感知信号，通过脑皮层视神经“V1 区→V2 区→V4 区→IT 区→AIT 区→PFC 区”的传输通道进行信息分析与过滤，具有视觉选择性和不变性双重特性。

5.3.4　知觉恒常性

知觉恒常性是人们知觉客观事物的一个重要的特性。

知觉该图时，我们会认为图 5.8(a)中上面的线比下面的线长，图 5.8(b)中上面的木头比下面的木头长，尽管两条线和两根木头长短一样。这是因为在现实的三维世界中，上面的线和木头会更长。

(a) 线段长度　　(b) 木头长度

图 5.8　庞佐错觉

大小恒常性(size constancy)即大小知觉恒常性。人对物体的知觉大小不完全随视像大小而变化，它趋向于保持物体的实际大小。大小知觉恒常性主要是过去经验的作用，例如，同一个人站在离我们 3m、5m、15m、30m 的不同距离处，他在我们视网膜上的视像随距离的不同而改变着(服从视角定律)。但是，我们看到这个人的大小却是不变的，仍然按他的实际大小来感知。例如，在图 5.8 中我们看到了庞佐错觉(Ponzo illusion)，图中央看起来大小不一的两个线条实际上是一样长的。庞佐错觉是因为两条趋近的线条造成了深度线索而产生的，不同深度的大小相同的图像通常显得大小不同。

形状恒常性(form constancy)即形状知觉恒常性。人从不同角度观察物体，或者物体位置发生变化时，物体在视网膜上的投射位置也发生了变化，但人仍然能够按照物体原来的形状来知觉(见图 5.9)。例如，房间门被打开时，它在视网膜上的视像形状与实际形状不完全一样，但看到门的形状仍是不变的。形状恒常性表明，物体的形状知觉具有相对稳定的特性。人的过往经验在形状恒常性中起重要作用。

图 5.9　形状知觉

颜色恒常性(color constancy)即颜色知觉恒常性。在不同的照明条件下，人们一般可正确地反映事物本身固有的颜色，而不受照明条件的影响。例如，不论在黄光还是在蓝光的照射下，人们总是把红旗知觉为红色的，而不是黄色的或是蓝色的。黑林认为，颜色知觉的恒常倾向是由于记忆色的影响。颜色恒常性可保证人对外界物体的稳定的辨认，具有明显的适应意义。

距离恒常性(distance constancy)又称距离的不变性,是指物体与知觉者的距离发生变化时,物体在网膜上造像的大小也发生相应的变化,但人知觉到的距离有保持原来距离的趋势的特性。

明度恒常性(brightness constancy)在不同照明条件下,人知觉到的明度不因物体的实际亮度的改变而变化,仍倾向于把物体的表面亮度知觉为不变。明度知觉恒常性是因人们考虑到整个环境的照明情况与视野内各物体反射率的差异,如果周围环境的亮度结构遭受不正常的变化,明度恒常性就会破坏。通常采用匹配法来研究明度常性,用邵勒斯比率来计算明度常性系数。

5.4 听觉感知

听觉过程包括机械→电→化学→神经冲动→中枢信息加工等环节。从外耳的集声至内耳基底膜的运动是机械运动,毛细胞受刺激后引起电变化,从而产生化学介质的释放、神经冲动的产生等活动。相关信息表传至中枢神经系统后,将发生一连串复杂的信息加工过程。

5.4.1 听觉通路

言语听觉比我们想象的要复杂得多,部分原因是口语速率最高达每秒 12 个音素(基本口语单位)。我们能理解的口语速度最多不能超过每分钟 50～60 个语音。在正常口语中,音素会出现重叠,同时存在一种协同发音现象,即一个语音片段的产生会影响后一个片段的产生,而线性问题是指协同发音引起言语知觉困难的现象。与线性问题相关的问题是非恒定性问题。这一问题是因任何给定的语音成分(如音素)的声音模式并不是恒定不变引起的,声音模式受到前后一个或多个语音成分的影响。这对辅音来说更是如此,因为它们的声音模式常常依赖于紧随其后的元音而定。

口语一般由连续变化的声音模式及少数停顿所组成。这与由独立声音构成的言语知觉形成鲜明对比。言语信号的连续性特征会产生分割问题,即决定一个连续的声音流怎样被分割成词汇。

从耳蜗到听觉皮质的听觉系统是所有感觉系统通路中最复杂的一种。听觉系统的每个水平上发生的信息过程和每一水平的活动都会影响较高水平和较低水平的活动。在听觉通路中,从脑的一边到另一边有广泛的交叉(见图 5.10)。

进入耳蜗神经核后,第八对脑神经听觉分支纤维终止于耳蜗核的背侧和腹侧。从两个耳蜗核分别发出纤维系统,从背侧耳蜗发出的纤维越过中线,然后经外侧丘系上升到皮质。外侧丘系最后终止于中脑的下丘,从腹侧耳蜗核发出的纤维,首先与同侧和对侧的上橄榄体复合体以突触联系,上橄榄体是听觉通路中的第一站,在这里发生两耳的相互作用。

上橄榄体复合体是听觉系统中令人感兴趣的中心,它由几个核组成,其中最大的是内侧上橄榄体和外侧上橄榄体。根据几种哺乳动物的比较研究,发现这两种核的大小与动物的感觉能力之间相互关联,Harrison 和 Irving 指出这两种核有不同的机能。他们指出,内侧上橄榄体和关联到眼球运动的声音定位有关,凡具有高度发展的视觉系统以及能注视声音的方向而做出反应的动物,内侧上橄榄核有着显著的外形。而另一方面,他们推论外侧上橄榄体则与独立于视觉系统以外的声音定位有关。具有敏锐的听觉但视觉能力有限的动物,都有显著的外侧上橄榄核。蝙蝠和海豚的视觉能力有限,但有极其发达的听觉系统,完全没有内侧上橄榄核。

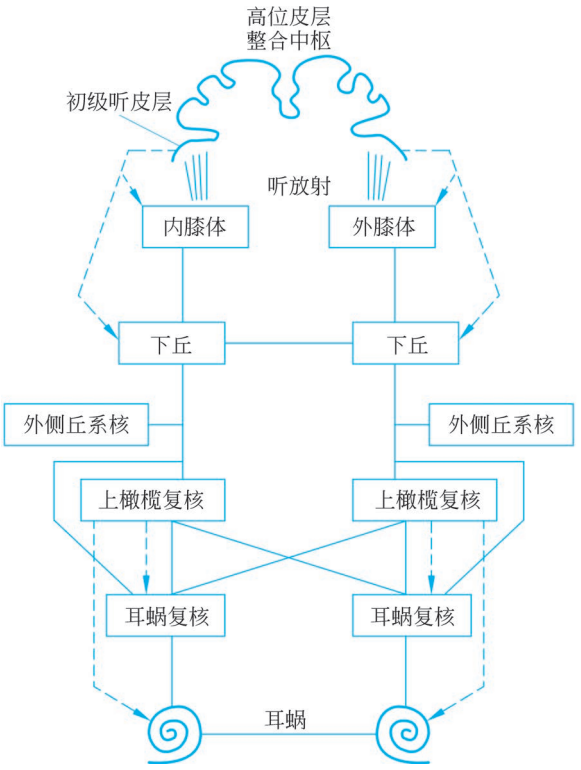

图 5.10　听觉通路

从上橄榄复合体出发的纤维上升经过外侧丘系到达下丘。从下丘系将冲动传达到丘脑的内侧膝状体。连接这两个区域的纤维束，叫作下丘臂。在内侧膝状体，听觉反射的纤维将冲动传导至颞上回(41 区和 42 区)，即听觉皮质区。

1988 年，伊里斯(A. W. Ellis)和杨(A. W. Young)提出了一个口语单词加工的模型(参见图 5.11)。这个模型包括以下 5 个成分。

(1) 听觉分析系统：用于从声波中提取音素和其他声音信息。

(2) 听觉输入词典：包含听者知道的关于口语单词的信息，但不包含语义信息。这个词典的目的就是通过恰当地激活词汇单元来识别熟悉单词。

(3) 语义系统：词义被存储于语义系统之中。

(4) 言语输出词典：用于提供单词的口语形式。

(5) 音素反应缓冲器：负责提供可分辨的口语声音。

这些成分可以各种方式组合起来，因此在听到一个单词至说出它之间存在 3 条不同的通路(见图 5.11)。

(1) 通路 1。这条通路利用听觉输入词典、语义系统和言语输出词典。它代表了无脑损伤人群正常识别和理解熟悉单词的认知通路。如果一个脑损伤患者只能利用这条通路(也许加上通路 2)，那么他将能够正确地说出熟悉单词。然而，在说出不熟悉单词或非词时将出现严重困难，因为这类材料没有存储于听觉输入词典之中。在这种情况下，患者需要使用通路 3。

(2) 通路 2。如果患者能够使用通路 2，但通路 1 和 3 受到严重损伤，那么他们应该能够重复熟悉的单词，但不能理解这些单词的意义。此外，患者也应该存在对非词的认知障碍，因为通路 2 不能处理非词信息。最后，由于这些患者将使用输入词典，所以他们应该能够区分词与非词。

(3) 通路 3。如果一个患者只损伤通路 3,那么他或她将展示在知觉和理解口语熟悉单词方面的完好的能力,但在知觉和重复不熟悉单词和非词时会出现障碍。这种情况临床上称为听觉性语音失认。然而,他阅读非词语时的能力完好。

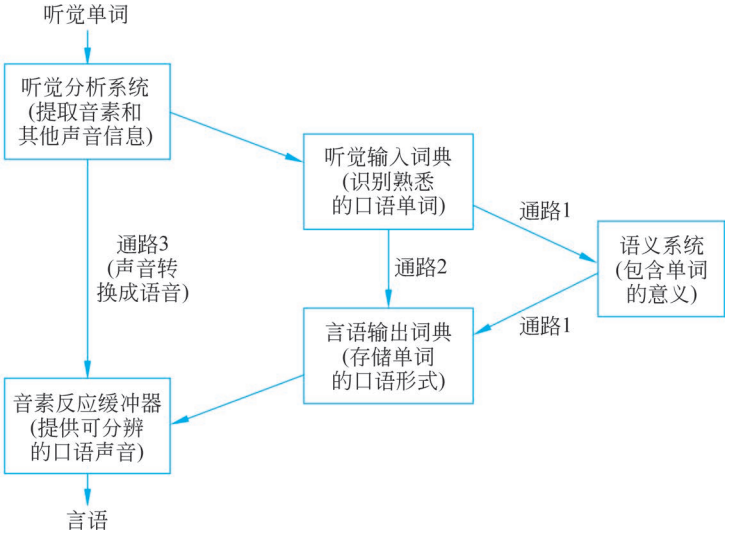

图 5.11 口语通路模型

5.4.2 语音编码

语音数字化的技术基本可以分为两大类:第一类方法是在尽可能遵循波形的前提下,将模拟波形进行数字化编码;第二类方法是对模拟波形进行一定处理,但仅对语音和收听过程中能够听到的语音进行编码。其中,语音编码的 3 种最常用的技术是脉冲编码调制(PCM)、差分 PCM(DPCM)和增量调制(DM)。通常,公共交换电话网中的数字电话都采用这 3 种技术。第二类语音数字化方法主要与用于窄带传输系统或有限容量的数字设备的语音编码器有关。采用该数字化技术的设备一般被称为声码器,声码器技术现在开始展开应用,特别是用于帧中继和 IP 上的语音。

除压缩编码技术外,人们还应用许多其他节省带宽的技术来减少语音所占带宽,优化网络资源。ATM 和帧中继网中的静音抑制技术可将连接中的静音数据消除,但并不影响其他信息数据的发送。语音活动检测(SAD)技术可以用来动态地跟踪噪声电平,并为这个噪声电平设置一个公用的语音检测阈值,这样就使得语音/静音检测器可以动态匹配用户的背景噪声环境,并将静音抑制的可听度降到最小。为了置换掉网络中的音频信号,这些信号不再穿过网络,舒适的背景声音在网络的任一端被集成到信道中,以确保话路两端的语音质量和自然声音的连接。语音编码方法归纳起来可以分成三大类:波形编码、信源编码、混合编码。

1. 波形编码

波形编码比较简单,编码前采样定理对模拟语音信号进行量化,然后进行幅度量化,再进行二进制编码。解码器作数/模变换后再由低通滤波器恢复出现原始的模拟语音波形,这就是最简单的脉冲编码调制(PCM),也称为线性 PCM。可以通过非线性量化,前后样值的差分、自适应预测等方法实现数据压缩。波形编码的目标是让解码器恢复出的模拟信号在波形上尽量与编码前原始波形相一致,也即失真要最小。波形编码的方法简单,数码率较高,在 64kb/s～32kb/s 时音质优良,当数码率低于 32kb/s 时音质明显降低,在 16kb/s 时音质非常差。

2. 信源编码

信源编码又称为声码器，是根据人声音的发声机理，在编码端对语音信号进行分析，分解成有声音和无声音两部分。声码器每隔一定时间分析一次语音，传送一次分析的编码有/无声和滤波参数。在解码端根据接收的参数再合成声音。声码器编码后的码率可以做得很低，如1.2kb/s、2.4kb/s，但是也有其缺点。首先是合成语音质量较差，往往清晰度可以而自然度没有，难以辨认说话人是谁，其次是复杂度比较高。

3. 混合编码

混合编码是将波形编码和声码器的原理结合起来，数码率在4kb/s ~16kb/s时音质比较好，最近有个别算法所取得的音质可与波形编码相当，复杂程度介乎于波形编码器和声码器之间。

上述的三大语音编码方案还可以分成许多不同的编码方案。语音编码属性可以分为4类，分别是比特速率、时延、复杂性和质量。比特速率是语音编码很重要的一方面。比特速率的范围可以是从保密的电话通信的2.4kb/s到64kb/s的G.711PCM编码和G.722宽带(7kHz)语音编码器。

5.4.3 语音识别

自动语音识别(automatic speech recognition，ASR)是实现人机交互尤为关键的技术，让计算机能够“听懂”人类的语音，将语音转换为文本。自动语音识别技术经过几十年的发展已经取得了显著的成效。近年来，越来越多的语音识别智能软件和应用走入了大家的日常生活，苹果的Siri、微软的小娜(Cortana)、百度度秘(Duer)、科大讯飞的语音输入法和灵犀等都是其中的典型代表。随着识别技术及计算机性能的不断进步，语音识别技术在未来社会中必将拥有更为广阔的前景。

1. 发展历程

以1952年贝尔实验室研制的特定说话人孤立词数字识别系统为起点，语音识别技术已经历了60多年的持续发展。其发展历程可大致分为以下4个阶段。

1) 20世纪50年代至70年代

该阶段是语音识别的初级阶段，主要研究孤立词识别。在动态时间规整技术、线性预测编码技术、矢量量化技术等取得进展。IBM公司的杰利内克(F. Jelinek)等在20世纪70年代末提出n-gram统计语言模型，并成功地将trigram模型应用于TANGORA语音识别系统中。此后美国卡内基梅隆大学采用bigram模型应用于SPHINX语音识别系统，大幅提高了识别率。此后一些著名的语音识别系统也相继采用bigram、trigram统计语言模型用于语音识别系统。

2) 20世纪80年代至90年代中期

识别算法从模式匹配技术转向基于统计模型的技术，更多地追求从整体统计的角度来建立最佳的话音识别系统。最典型的为隐马尔可夫模型(hidden Markov model，HMM)在大词汇量连续语音识别系统中的成功应用。美国国防部先进研究项目局(defense advanced research projects agency，DARPA)自1983年开始为期10年的DARPA战略计算工程项目，其中包括用于军事领域的语音识别和语言理解、通用语料库等。参加单位包括MIT(麻省理工学院)、CMU(卡内基梅隆大学)、BellLab和IBM公司等。20世纪80年代末，美国卡耐基梅隆大学用VQ-HMM实现了语音识别系统SPHINX，这是世界上第一个高性能的非特定人、

大词汇量、连续语音识别系统，开创了语音识别的新时代。至 90 年代中期，语音识别技术进一步成熟，并出现了一些很好的产品。该阶段可以认为是统计语音识别技术的快速发展阶段。

3) 20 世纪 90 年代中期至 21 世纪初

该阶段语音识别研究工作更趋于解决在真实环境应用时所面临的实际问题。美国国家标准技术局和美国国防部先进研究项目局组织了大量的语音识别技术评测，极大地推动了该技术的发展。在此阶段，基于高斯混合模型(Gaussian mixture model，GMM)和 HMM 的混合语音识别框架成为领域内主流技术。而区分度训练技术的提出，进一步提升了系统性能。此外，为提升系统的鲁棒性及实用性，语音抗噪技术、说话人自适应训练(speaker adaptive training，SAT)等技术被相继提出。该阶段可看作 GMM-HMM 混合语音识别技术趋于成熟并应用的阶段。

4) 21 世纪初至今

该阶段的特点是基于深度学习的语音识别技术成为主流，以 2011 年提出的上下文相关-深度神经网络-隐马尔可夫框架为变革开始的标志。基于链接时序分类(connectionist temporal classification，CTC)搭建过程简单，且在某些情况下性能更好。2016 年，谷歌公司提出 CD-CTC-SMBR-LSTM-RNNS，标志着传统的 GMM-HMM 框架被完全替代。声学建模由传统的基于短时平稳假设的分段建模方法变革到基于不定长序列的直接判别式区分的建模。由此，语音识别性能逐渐接近实用水平，而移动互联网的发展同时带来了对语音识别技术的巨大需求，两者相互促进。与深度学习相关的参数学习算法、模型结构、并行训练平台等成为该阶段的研究热点。该阶段可看作深度学习语音识别技术高速发展并大规模应用的阶段。

我国语音识别研究工作起步于 20 世纪 50 年代，而研究热潮是从 20 世纪 80 年代中期开始。在 863 计划的支持下，中国开始了有组织的语音识别技术的研究。语音识别正逐步成为信息技术中人机接口的关键技术，研究水平也从实验室逐步走向实用。

2. 语音识别系统结构

语音识别系统包含 4 个主要模块：信号处理、解码器、声学模型、语言模型(见图 5.12)。

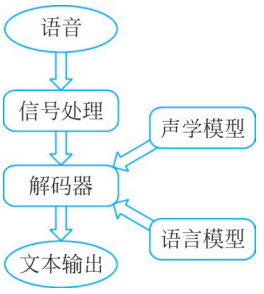

图 5.12 语音识别系统框架

信号处理模块输入为语音信号，输出为特征向量，随着远场语音交互需求越来越大，前端信号处理与特征提取在语音识别中位置越来越重要。一般而言，主要过程为首先通过麦克风阵列进行声源定位，然后消除噪声。通过自动增益控制将收音器采集到的声音放到正常幅值。通过去噪等方法对语音进行增强，然后将信号由时域转换到频域，最后提取适用于 AM 建模的特征向量。

声学模型对声学和发音学知识进行建模，其输入为特征抽取模块产生的特征向量，输出为某条语音的声学模型得分。声学模型是对声学、语音学、环境的变量，以及说话人性别、口音的差异等的知识表示。声学模型的好坏直接决定整个语音识别系统的性能。

语言模型则是对一组字序列构成的知识表示，用于估计某条文本语句产生的概率，称为语言模型得分。模型中存储的是不同单词之间的共现概率，一般通过从文本格式的语料库中估计得到。语言模型与应用领域和任务密切相关，当这些信息已知时，语言模型得分更加精确。

解码器根据声学模型和语言模型，将输入的语音特征矢量序列转换为字符序列。解码器将所有候选句子的声学模型得分和语言模型得分融合在一起，输出得分最高的句子作为最终

的识别结果。

3. 基于深度神经网络的语音识别系统

基于深度神经网络的语音识别系统框架如图 5.13 所示。相比于传统的基于 GMM-HMM 的语音识别系统，其最大的改变是采用深度神经网络替换 GMM 模型对语音的观察概率进行建模。最初主流的深度神经网络是最简单的前馈型深度神经网络(feedforward deep neural network，FDNN)。DNN 相比 GMM 的优势在于：①使用 DNN 估计 HMM 的状态的后验概率分布不需要对语音数据分布进行假设；②DNN 的输入特征可以是多种特征的融合，包括离散或者连续的；③DNN 可以利用相邻的语音帧所包含的结构信息。

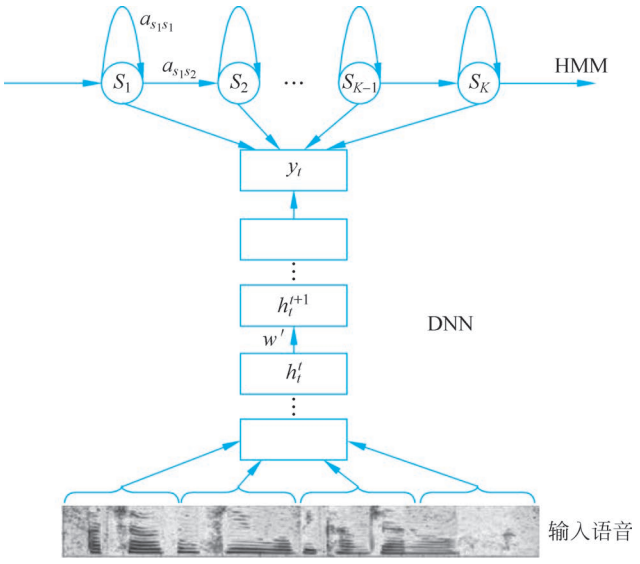

图 5.13 基于深度神经网络的语音识别系统框架

考虑到语音信号的长时相关性，一个自然而然的想法是选用具有更强长时建模能力的神经网络模型。于是，循环神经网络(recurrent neural network，RNN)近年来逐渐替代传统的 DNN 成为主流的语音识别建模方案。如图 5.14 所示，相比于前馈型神经网络(DNN)，循环神经网络在隐层上增加了一个反馈连接，也就是说，RNN 隐层当前时刻的输入有一部分是前一时刻的隐层输出，这使得 RNN 可以通过循环反馈连接看到前面所有时刻的信息，这赋予了 RNN 记忆功能。这些特点使得 RNN 非常适合用于对时序信号的建模。而长短时记忆模块(long-short term memory，LSTM)的引入解决了传统简单 RNN 梯度消失等问题，使得 RNN 框架可以在语音识别领域实用化并获得了超越 DNN 的效果，目前已经使用在业界一些比较

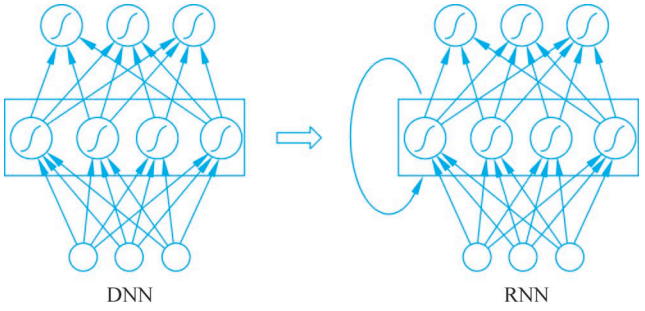

图 5.14 DNN 和 RNN 示意图

先进的语音系统中。除此之外，研究人员还在 RNN 的基础上做了进一步改进工作，如图 5.15 是当前语音识别中的主流 RNN 声学模型框架，主要包含两部分：深层双向 RNN 和链接时序分类 CTC 输出层。其中双向 RNN 对当前语音帧进行判断时，不仅可以利用历史的语音信息，还可以利用未来的语音信息，从而进行更加准确的决策；CTC 使得训练过程无需帧级别的标注，实现有效的“端对端”训练。

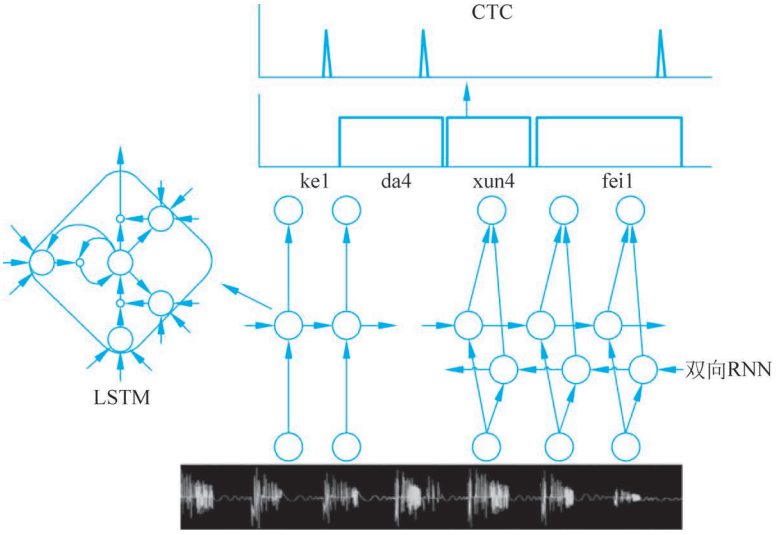

图 5.15　基于 RNN-CTC 的主流语音识别系统框架

语音识别任务是将输入波形映射到最终的词序列或中间的音素序列。声学模型真正应该关心的是输出的词或音素序列，而不是在传统的交叉熵训练中优化、一帧一帧地标注。为了应用这种观点并将语音输入帧映射成输出标签序列，链接时序分类 CTC 方法被引入了进来。为了解决语音识别任务中输出标签数量少于输入语音帧数量的问题，链接时序分类 CTC 引入了一种特殊的空白标签，并且允许标签重复，从而迫使输出和输入序列的长度相同。

链接时序分类 CTC 的一个特点是我们可以选择大于音素的输出单元，如音节和词。这说明输入特征可以使用大于 10ms 的采样率构建。链接时序分类 CTC 提供了一种以端到端的方式优化声学模型的途径。用端到端的语音识别系统直接预测字符而非音素，从而也就不再需要使用的词典和决策树了。

早在 2012 年，卷积神经网络 CNN 就被用于语音识别系统，但始终没有明显的突破。最主要的原因是没有突破传统前馈神经网络采用固定长度的帧拼接作为输入的思维定式，从而无法看到足够长的语音上下文信息。另外一个缺陷是只将 CNN 视作一种特征提取器，因此所用的卷积层数很少，一般只有一到两层，这样的卷积网络表达能力十分有限。

讯飞研发了深度全序列卷积神经网络(deep fully convolutional neural network，DFCNN)的语音识别框架，使用大量的卷积层直接对整句语音信号进行建模，更好地表达了语音的长时相关性。DFCNN 的结构如图 5.16 所示，它直接将一句语音转换成一张图像作为输入，即先对每帧语音进行傅里叶变换，再将时间和频率作为图像的两个维度，然后通过非常多的卷积层和池化(pooling)层的组合，对整句语音进行建模，输出单元直接与最终的识别结果，如音节或者汉字相对应。

DFCNN 直接将语谱图作为输入，与其他以传统语音特征作为输入的语音识别框架相比具有天然的优势。从模型结构来看，DFCNN 与传统语音识别中的 CNN 做法不同，它借鉴了

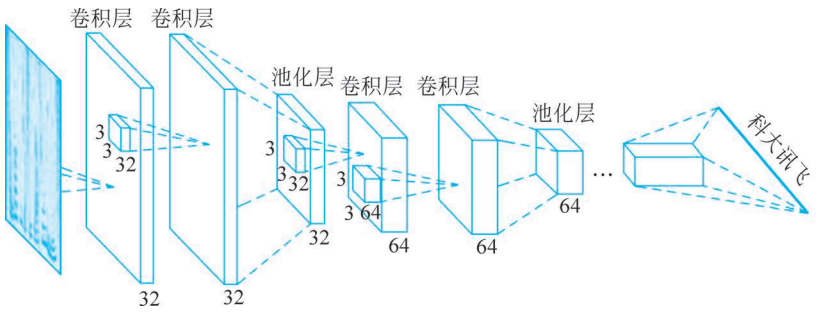

图 5.16 DFCNN 示意图

图像识别中效果最好的网络配置，每个卷积层使用 3×3 的小卷积核，并在多个卷积层之后再加上池化层，这样大大增强了 CNN 的表达能力，与此同时，通过累积非常多的这种卷积池化层对，DFCNN 可以看到非常长的历史和未来信息，这就保证了 DFCNN 可以出色地表达语音的长时相关性，相比 RNN 网络结构在鲁棒性上更加出色。最后，从输出端来看，DFCNN 还可以和近期很热的 CTC 方案完美结合以实现整个模型的端到端训练，且其包含的池化层等特殊结构可以使得以上端到端训练变得更加稳定。

在和其他多个技术点结合后，科大讯飞 DFCNN 的语音识别框架在内部数千小时的中文语音短信听写任务上，相比目前业界最好的语音识别框架双向 RNN-CTC 系统获得了 15%的性能提升，同时结合科大讯飞的 HPC 平台和多 GPU 并行加速技术，训练速度也优于传统的双向 RNN-CTC 系统。DFCNN 的提出开辟了语音识别的一片新天地。

5.4.4 语音合成

语音合成即让计算机生成语音的技术，其目标是让计算机能输出清晰、自然、流畅的语音。按照人类言语功能的不同层次，语音合成也可以分成 3 个层次，即从文字到语音的合成、从概念到语音的合成、从意向到语音的合成。这 3 个层次反映了人类大脑中形成说话内容的不同过程，涉及人类大脑的高级神经活动。目前成熟的语音合成技术只能够完成从文字到语音(text-to-speech，TTS)的合成，该技术也常常被称作文语转换技术。

典型的文字到语音合成系统如图 5.17 所示，该系统可以分为文本分析模块、韵律预测模块和声学模型模块，下面对 3 个模块进行简要的介绍。

图 5.17 典型的文字到语音合成系统结构图

1. 文本分析模块

文本分析模块是语音合成系统的前端。它的作用是对输入的任意自然语言文本进行分析，输出尽可能多语言相关的特征和信息，为后续的系统提供必要的信息。它的处理流程依次为：文本预处理、文本规范化、自动分词、词性标注、字音转换、多音字消歧、字形到音素(grapheme to phoneme，G2P)、短语分析等。文本预处理包括删除无效符号、断句等。其中，文本规范化的任务就是将文本中的非普通文字(如数学符号、物理符号等)字符识别出来，并转换为一种规范化的表达。字音转换的任务是将待合成的文字序列转换为对应的拼音序列。多音字消歧则是解决一字多音的问题。G2P 是为了处理文本中可能出现的未知读音的字词，这在英文或其他单词以字母组成的语言中经常出现。

2. 韵律预测模块

韵律即是实际话流中的抑扬顿挫和轻重缓急，例如重音的位置分布及其等级差异，韵律边界的位置分布及其等级差异，语调的基本骨架及其跟声调、节奏和重音的关系等。由于这些特征需要通过不止一个音段上的特征变化得以实现，通常也称之为超音段特征。韵律表现是一个很复杂的现象，对韵律的研究涉及语音学、语言学、声学、心理学等多个领域。韵律预测模块则接收文本分析模块的处理结果，预测相应的韵律特征，包括停顿、句重音等超音段特征。韵律模块的主要作用是保证合成语音拥有自然的抑扬顿挫，提高语音的自然度。

3. 声学模型模块

声学模型的输入为文本分析模块提供的文本相关特征和韵律预测模块提供的韵律特征，输出为自然语音波形。目前主流的声学模型采用的方法可以概括为两种：一种是基于时域波形的拼接合成方法，声学模型模块首先对基频、时长、能量和节奏等信息建模，并在大规模语料库中根据这些信息挑选最合适的语音单元，然后通过拼接算法生成自然语音波形；另一种是基于语音参数的合成方法，声学模型模块根据韵律和文本信息的指导来得到语音的声学参数，如谱参数、基频等，然后通过语音参数合成器来生成自然语音波形。

语音合成系统的声学模型从所采用的基本策略来看，可以分为基于发音器官的模型和基于信号的模型两大类。前者试图对人类的整个发音器官进行直接建模，通过该模型进行语音的合成，该方法也被称为基于生理参数的语音合成。后者则是基于语音信号本身进行建模或者直接进行基元选取拼接合成。相比较而言，基于信号模型的方法具有更强的应用价值，因而得到了更多研究者和工业界的关注。基于信号模型的方法有很多，主要包括基于基元选取的拼接合成和统计参数语音合成。

5.5 人脸识别

人脸识别技术是指利用分析比较的计算机技术识别人脸。人脸识别技术是基于人的脸部特征，对输入的人脸图像或者视频流，首先判断其是否存在人脸，如果存在人脸，则进一步给出每个脸的位置、大小和各主要面部器官的位置信息，并依据这些信息，进一步提取每个人脸中所蕴含的身份特征，并将其与已知的人脸进行对比，从而识别每个人脸的身份。

人脸识别技术识别过程一般分三步。

(1) 首先建立人脸的面像档案。用摄像机采集单位人员的人脸的面像文件或取他们的照片形成面像文件，并将这些面像文件生成面纹(Faceprint)编码存储起来。

(2) 获取当前的人体面像。用摄像机捕捉到当前出入人员的面像，或取照片输入，并将当前的面像文件生成面纹编码。

(3) 用当前的面纹编码与档案库存的面纹编码比对。将当前的面像的面纹编码与档案库存中的面纹编码进行检索比对。上述的“面纹编码”方式是根据人脸脸部的本质特征工作的。这种面纹编码可以抵抗光线、皮肤色调、面部毛发、发型、眼睛、表情和姿态的变化，具有强大的可靠性，从而使它可以从百万人中精确地辨认出某个人。人脸的识别过程，利用普通的图像处理设备就能自动、连续、实时地完成。

人脸识别系统主要包括 4 个组成部分，即人脸图像采集(图的左侧)、人脸预处理、特征提取、特征比对(见图 5.18)。

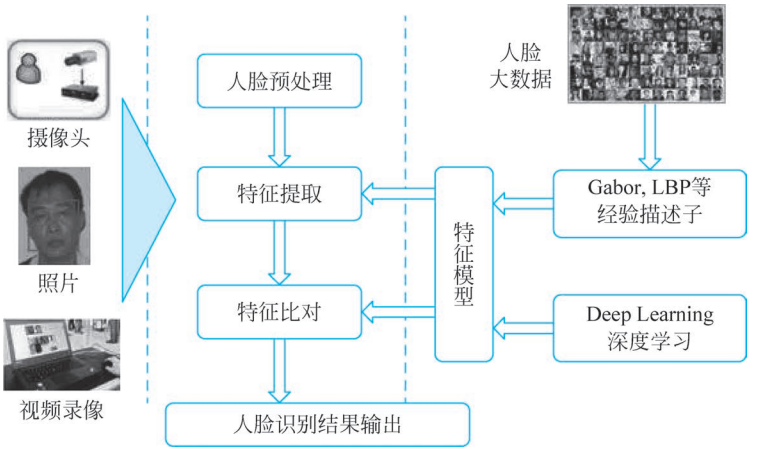

图 5.18　人脸识别系统

1. 人脸图像采集及检测

人脸图像采集：不同的人脸图像都能通过摄像镜头采集下来，例如静态图像、动态图像、不同的位置、不同表情等方面都可以得到很好的采集。当用户在采集设备的拍摄范围内时，采集设备会自动搜索并拍摄用户的人脸图像。

人脸检测：人脸检测在实际中主要用于人脸识别的预处理，即在图像中准确标定出人脸的位置和大小。人脸图像中包含的模式特征十分丰富，如直方图特征、颜色特征、模板特征、结构特征及 Haar 特征等。人脸检测就是把其中有用的信息挑出来，并利用这些特征实现人脸检测。

主流的人脸检测方法基于以上特征采用 Adaboost 学习算法，Adaboost 算法是一种用来分类的方法，它把一些比较弱的分类方法合在一起，组合出新的很强的分类方法。

人脸检测过程中使用 Adaboost 算法挑选出一些最能代表人脸的矩形特征（弱分类器），按照加权投票的方式将弱分类器构造为一个强分类器，再将训练得到的若干强分类器串联组成一个级联结构的层叠分类器，有效地提高分类器的检测速度。

2. 人脸图像预处理

对于人脸的图像预处理是基于人脸检测结果，对图像进行处理并最终服务于特征提取的过程。系统获取的原始图像由于受到各种条件的限制和随机干扰，往往不能直接使用，必须在图像处理的早期阶段对它进行灰度校正、噪声过滤等图像预处理。对于人脸图像而言，其预处理过程主要包括人脸图像的光线补偿、灰度变换、直方图均衡化、归一化、几何校正、滤波以及锐化等。

3. 人脸图像特征提取

人脸识别系统可使用的特征通常分为视觉特征、像素统计特征、人脸图像变换系数特征、人脸图像代数特征等。人脸特征提取就是针对人脸的某些特征进行的。人脸特征提取，也称人脸表征，它是对人脸进行特征建模的过程。人脸特征提取的方法归纳起来分为两大类：一种是基于知识的表征方法；另一种是基于代数特征或统计学习的表征方法。

基于知识的表征方法主要是根据人脸器官的形状描述以及它们之间的距离特性来获得有助于人脸分类的特征数据，其特征分量通常包括特征点间的欧氏距离、曲率和角度等。人脸由眼睛、鼻子、嘴、下巴等局部构成，对这些局部和它们之间结构关系的几何描述，可作为识别人

脸的重要特征，这些特征被称为几何特征。基于知识的人脸表征主要包括基于几何特征的方法和模板匹配法。

4. 人脸图像匹配与识别

提取的人脸图像的特征数据与数据库中存储的特征模板进行搜索匹配，通过设定一个阈值，当相似度超过这一阈值，则把匹配得到的结果输出。人脸识别就是将待识别的人脸特征与已得到的人脸特征模板进行比对，根据所提取特征的相似程度对人脸的身份信息进行判断。这一过程又分为两类：一类是确认，是一对一地进行图像比较的过程；另一类是辨认，是一对多地进行图像匹配对比的过程。

人脸识别技术广泛用于政府、军队、银行、社会福利保障、电子商务、安全防务等领域。例如电子护照及身份证，这是规模最大的应用。在国际民航组织(ICAO)已确定，从 2010 年 4 月 1 日起，在其 118 个成员国家和地区，人脸识别技术是首推识别模式，该规定已经成为国际标准。美国已经要求和它有出入免签证协议的国家在 2006 年 10 月 26 日之前必须使用结合了人脸指纹等生物特征的电子护照系统，到 2006 年年底已经有 50 多个国家实现了这样的系统。