# 第一章　概述



## 1.1 什么是计算机视觉

计算机视觉既是工程领域、也是科学领域中的一个富有挑战性的重要研究领域。计算机视觉是一门综合性的学科，它已经吸引了来自各个学科的研究者参加到对它的研究之中，其中包括计算机科学和工程、信号处理、物理学、应用数学和统计学、神经生理学和认知科学等。

视觉是各个应用领域，如制造业、检验、文档分析、医疗诊断和军事等领域中各种智能／自主系统中不可分割的一部分。由于它的重要性，一些先进国家，例如美国把对计算机视觉的研究列为对经济和科学有广泛影响的科学和工程中的重大基本问题，即所谓的重大挑战（grand challenge）。“计算机视觉的挑战是要为计算机和机器人开发具有与人类水平相当的视觉能力。机器视觉需要图象信号，纹理和颜色建模，几何处理和推理，以及物体建模。一个有能力的视觉系统应该把所有这些处理都紧密地集成在一起。”[Neg91]作为一门学科，计算机视觉开始于 60 年代初，但在计算机视觉的基本研究中的许多重要进展是在 80 年代取得的。现在计算机视觉已成为一门不同于人工智能、图象处理、模式识别等相关领域的成熟学科。计算机视觉与人类视觉密切相关，对人类视觉有一个正确的认识将对计算机视觉的研究非常有益。为此我们将先介绍人类视觉。

### 1.1.1 人类视觉

感觉是人的大脑与周围世界联系的窗口，它的任务是识别周围的物体，并判断这些物体之间的关系。我们的思维活动是以我们对客观世界与环境的认识为基础的，而感觉则是客观世界与我们对环境的认识之间的桥梁，使我们的思维与周围世界建立某种对应关系。视觉则是人类最重要的感觉，它是人的主要感觉来源，人类认识外界信息中 80%来自视觉。

人有多种感觉，但对人的智力产生影响的主要是视觉和听觉。味觉和嗅觉是丰富多样的，

但很少有人去思考它们。在视觉和听觉中形状、色彩、运动、声音等很容易就被结合成各种明确和高度复杂、多样的空间和时间的组织结构。所以这两种感觉就成了理智活动得以行使和发挥作用的非常合适的媒介和环境。但人听到的声音要想具有意义还需要联系其它的感性材料。而视觉则不同，它是一种高度清晰的媒介，它提供关于外界世界中各种物体和事件的丰富信息。因此它是思维的一种最基本的工具。

视觉对正常人来说是生而有之，毫不费力的能力。但实际上视觉系统所完成的功能却是十分复杂的。有人认为视觉本身就包含了思维的一切基本因素。设想你要在一个会场中寻找一位朋友，呈现在你眼前的是由参加会议的人、桌、椅、主席台等组成的复杂景物。眼睛得到这些信息以后先要对景物的各部分进行分类，然后从中选出与朋友的外表有关的特征作出判断，那么在人的眼睛视网膜上映照的景物成象是否就能直接提供判断时所需要的有关特征呢？不是的，这里需要大脑的思考。例如，虽然人在不同距离处观察同一物体时在眼睛中成象的大小是不同的。但人们在观察某人以便估计他的身高时却不会因为他在近处而感到他高些，也不会因他在远处而感到他矮些。这是大脑根据被观察物体的距离与周围物体的比较，并依靠有关的知识对输入的图象信息进行处理、解释的结果。如果你是在一个灯光暗淡的剧院中寻找朋友，这个问题就变得更为困难。你刚走进剧院时会感到一片漆黑看不清东西，过了几分钟你的眼睛变得习惯于在黑暗中观察。事实上你的视觉系统在此期间中对微光变得更敏感了。但这时许多本来可用的信息丧失了，物体可能难以与背景相区分，许多细节难以分辨。即使这样人也总能认出朋友。总之，视觉是一个复杂的感知和思维过程，视觉器官一眼睛接受外界的刺激信息，而大脑对这些信息通过复杂的机理进行处理和解释，使这些刺激具有明确的物理意义。

从以上分析我们还可以看到敏感（Sansation）、感觉（Perception）、认知（Cognition）这三个概念之间的联系和差别。敏感是把外界的各种刺激转换成人体神经系统能够接受的生物电信号。它所完成的是信号的转换，并不涉及对信号的理解。例如，人眼是视觉的敏感器官，它使光信号通过视网膜转换电信号。与摄象机的光电传感器相似，视网膜的感光细胞对光信号在平面上进行采样，产生点阵形式的电信号，所不同的仅是摄象机的空间采样是均匀的，而视网膜的采样是不均匀的，在中央凹附近采样分辨率高，而在周围的分辨率低。而感觉的任务是把敏感器官的各种输入转换和处理成为对外部世界的理解。例如，对视觉来说就是能说出周围世界中有什么东西和这些东西之间的空间关系。这些都是关于周围世界的概念。从输入的点阵形式的信号到形式化的对客观世界的各种概念其中要经过复杂的信息处理和推理。而认知是以人们对周围客观世界的概念为基础的。如果没有感觉这个人与外部世界的桥梁或窗口，人的思维活动就失去基本的依据。

## 1.1.2 计算机视觉

人类正在进入信息时代，计算机将越来越广泛地进入几乎所有领域。一方面是更多未经计算机专业训练的人也需要应用计算机，而另一方面是计算机的功能越来越强，使用方法越来越复杂。这就使人在进行交谈和通讯时的灵活性与目前在使用计算机时所要求的严格和死板的方式之间产生了尖锐的矛盾。人可通过视觉、听觉和语言与外界交换信息，并且可用不同的方式表示相同的含义，而目前的计算机却要求严格按照各种程序语言来编写程序，只有这样计算机才能运行。为使更多的人能使用复杂的计算机，必须改变过去的那种让人来适应计算机、死记硬背计算机使用规则的情况。而是反过来让计算机来适应人的习惯和要求，以人所习惯的方式与人进行信息交换，也就是让计算机具有视觉、听觉和说话等能力。这时计算机必须具有逻辑推理和决策的能力。具有上述能力的计算机就是智能计算机。

智能计算机不仅使计算机更便于为人们所使用，同时如果用这样的计算机来控制各种自

动化装置特别是智能机器人，就可以使这些自动化系统和智能机器人具有适应环境和自主作出决策的能力。这就可以在各种场合取代人的繁重工作，或代替人到各种危险和恶劣环境中完成任务。

计算机视觉就是用各种成象系统代替视觉器官作为输入敏感手段，由计算机来代替大脑完成处理和解释。计算机视觉的最终研究目标就是使计算机能象人那样通过视觉观察和理解世界，具有自主适应环境的能力。这是要经过长期努力才能达到的目标。因此，在实现最终目标以前，人们努力的中期目标是建立一种视觉系统，这个系统能依据视觉敏感和反馈以某种程度的智能完成一定的任务。例如，计算机视觉的一个重要应用领域就是自主车辆的视觉导航，目前还没有条件实现象人那样识别和理解任何环境、完成自主导航的系统。因此，目前人们努力的研究目标是实现在高速公路上具有道路跟踪能力，可避免与前方车辆碰撞的视觉辅助驾驶系统。这里要指出的一点是在计算机视觉系统中计算机起代替人脑的作用，但并不意味着计算机必须按人类视觉的方法完成视觉信息的处理。计算机视觉可以而且应该根据计算机系统的特点来进行视觉信息的处理。但是，人类视觉系统是迄今为止，人们所知道的功能最强大和完善的视觉系统。在以下的章节中我们会看到，对人类视觉处理机制的研究将给计算机视觉的研究提供启发和指导。因此，用计算机信息处理的方法研究人类视觉的机理，建立人类视觉的计算理论，也是一个非常重要和令人感兴趣的研究领域。这方面的研究被称为计算视觉（Computational Vision）。计算视觉可被认为是计算机视觉中的一个研究领域。

有不少学科的研究目标与计算机视觉相近或与此有关。这些学科包括图象处理、模式识别或图象识别、景物分析、图象理解等。由于历史发展或领域本身的特点这些学科互有差别，但又有某种程度的相互重迭。为了清晰起见，我们把这些与计算机视觉有关的学科从研究目标和方法的角度加以归纳。

**1. 图象处理**

图象处理技术把输入图象转换成具有所希望特性的另一幅图象。例如，可通过处理使输出图象有较高的信噪比，或通过增强处理突出图象的细节，以便于操作员的检验。在计算机视觉研究中经常利用图象处理技术进行预处理和特征抽取。

**2. 模式识别（图象识别）**

模式识别技术根据从图象抽取的统计特性或结构信息，把图象分成预定的类别。例如，文字识别或指纹识别。在计算机视觉中模式识别技术经常用于对图象中的某些部分（例如分割区域）的识别和分类。

**3. 图象理解（景物分析）**

给定一幅图象，图象理解程序不仅描述图象本身，而且描述和解释图象所代表的景物，以便对图象代表的内容作出决定。在人工智能视觉研究的初期经常使用景物分析这个术语，以强调二维图象与三维景物之间的区别。图象理解除了需要复杂的图象处理以外还需要具有关于景物成象的物理规律的知识以及与景物内容有关的知识。

在建立计算机视觉系统时需要用到上述学科中的有关技术，但计算机视觉研究的内容要比这些学科更为广泛。计算机视觉的研究与人类视觉的研究密切相关（见 1.3.5 中的论述）。为实现建立与人的视觉系统相类似的通用计算机视觉系统的目标需要建立人类视觉的计算理论。

## 1.2 视觉的计算理论

视觉是一个根据图象发现周围景物中有什么物体和物体在什么地方的过程，也就是从图象得到对观察者有用的符号描述的过程。因此，视觉是一个有明确输入和输出的信息处理问题。

对计算机视觉系统来说，输入是表示三维景物投影的灰度阵列。可以有若干个输入阵列，这些阵列可提供不同方面或不同视角，不同时刻或在不同波长得到的信息。希望的输出是对图象所代表景物的符号描述。这些描述的确切本质取决于观察的目标和期望。通常这些描述是关于物体的类别和物体间的关系，但也可能包括如表面空间结构，表面物理特性（形状、纹理、颜色、材料、阴影）以及光源位置这样的信息。

从输入图象到得出景物描述之间存在着巨大的间隙，需要经过一系列的信息处理和理解过程。对这个过程本质的认识是揭开视觉之谜的关键，但目前我们对这些还远未了解清楚。以下我们对此过程作初步的分析。通过视觉识别物体就是把图象的元素与已知的景物中的物体的描述或模型之间建立对应关系。图象中的元素是点状的象素，象素的值就是这个象素处的灰度值，这是点状的数据。而与此相对，物体是通过它的形状、大小、几何结构、颜色等特征来描述的。这些特征代表物体的整体性质。要在输入的点状数据与物体的整体性质之间建立对应关系就必须要经过一个把点状数据聚集（grouping）起来的过程。这样的聚集过程不只是在视觉中有，而且在听觉及其它感觉中也存在。

与如何形成整体性质相联系的问题是恒常性问题。大家都知道，图象中各点的灰度是景物中多种因素综合作用的结果。这些因素包括光照条件、物体表面的反射特性、观察者相对于物体的距离和方位、物体表面形状等。这些因素的任何变化都会改变图象的灰度，也就会改变我们看到的图象。但是我们通过视觉所感觉到的物体的形状、大小和颜色都是与观察者的状况以及照明条件无关的。具体而言，当照明条件和观察者相对物体的距离方位发生变化时，虽然在视网膜上产生的图象要随之而变化，但人看到的总是某种形状和大小的物体。例如，当你从不同角度和距离观察一张桌子时，桌子在你的眼睛视网膜上的成像会随之而改变，但你看到的始终是一定大小和形状的桌子。外部世界投影在视网膜上产生了图象，这是一个敏感的过程。这个过程得到的图象是以点的方式组织在一起的，是经常变化的。但人在大脑中感觉到的是物体可变的外表后面的恒定特征。因此，大脑不但把点状的传感信息聚集成整体，而且经过一个因素分解过程（factoring）把这些影响传感器信息的条件，即照明条件、观察者的距离和方位等因素分离出去，得到纯粹的关于物体的信息。这些信息是不随上述条件而变的，因此被称为恒常性（constancies）。总之，大脑不是直接根据外部世界在视网膜上的投影成象，而是根据经过聚集过程和因素分解过程处理以后的信息来识别物体的[Kan 87]。

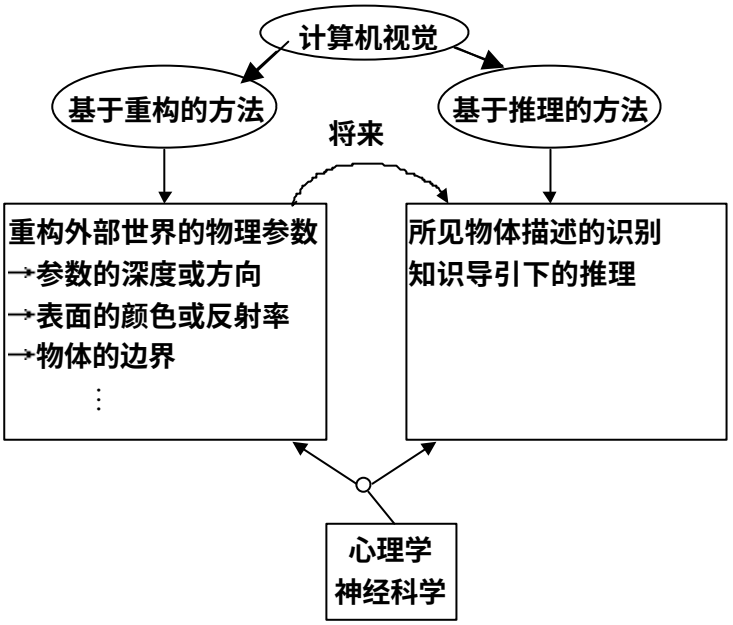


图 1.1　计算机视觉中的两种方法

与分析上述两种处理过程有关的一个非常重要的问题，是把点状的图象信息变换成整体描述的聚集过程与对各种影响成象结果的因素进行分解的因素分解过程之间的关系。在没有完成因素分解过程以前我们能着手进行聚集过程吗？以 Marr[Mar 82]为首的一些科学家认为在得到关于物体的纯净的信息（clean information），例如深度、表面、方向、反射率等以前，

做任何聚集的处理都是无用的。他们把这样的纯净信息称为本征图象（intrinsic image），因此他们采用基于重构（reconstruction）的视觉信息处理方法，也就是通过重构这些本征图象来识别物体。而另一派科学家则认为某些预先进行的聚集过程不仅可以为因素分解过程提供必要的基础，而且还可形成某种反应物体空间结构的图象关系，根据这些图象关系可以产生对图象内容的假设。因此，他们采用了基于推理和识别的视觉信息处理方法。前一种观点是以 Marr 关于人类视觉的计算理论为代表；后一种观点是以 Gestalt（Gestalt）学派，及其后续者，如 Lowe[Low 85]、Pentland[Pen 88]等关于感知组织（Perception organization）的理论为代表。这两派理论各自反映了视觉过程中的基本矛盾，但都未能对视觉过程作出满意的解释。这两种理论的争论推动了对视觉的研究（见图 1.1）。

## 1.2.1 Marr[Mar 82]（马尔）的视觉计算理论

Marr 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。Marr 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。

Marr 认为视觉是一个信息处理过程。这个过程根据外部世界的图象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一步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象。然而，数据进入我们的眼睛是要以光线为媒介的。灰度图象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 Marr 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象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）简图或本征图象。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象这个中间表象层次。因为这个中间表象已经从原始的图象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。根据本征图象表示的这些信息可以可靠地把图象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）。Marr 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。

Marr 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。对此现在已有不少学者提出改进意见，关于这个问题将在第二章中详细讨论。

## 1.2.2 基于推理的视觉理论

由于只根据图象数据本身不能对相应的物体空间结构提供充分的约束，也就是说这是一个约束不充分（underconstrained）的问题。因此，为了理解图象的内容必须要有附加的约束条件。Gestalt 心理学家发现的感知组织现象是一种非常有力的关于象素整体性的附加约束。

从而为视觉推理提供了基础。Gestalt 是德文 Gestalt 的译音。英文中常译成 form（形式）或 shape（形状）。Gestalt 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。在视觉研究中 Gestalt 理论认为把点状数据聚集成整体特征的聚集过程是所有其它有意义的处理过程的基础。人的视觉系统具有在对景物中的物体一无所知的情况下从景物的图象中得到相对的聚集（grouping）和结构的能力。这种能力被称为感知组织。按 Gestalt 理论感知组织的基本原理被称为 Pragmant，意即“简约合宜”。它来源于 Gestalt 心理学家发现有些“形”给人的的感受是极为愉悦的。这就是那些在特定条件下视觉刺激被组织得最好、最规则（对称、统一、和谐）、具有最大限度的简单明了性的“形”。对这种形他们发明了一个独特的字眼，即 Pragnant，有人把这个词译成“完形”。人的视觉系统具有很强的检测多种图案和随机的、但又有显著特色的图象元素排列的能力。例如，人可从随机分布的图象元素中立即检测出对称性、集群、共线性、平行性、连通性和重复纹理等。感知组织把点状的传感数据变换成客观的表象。在这些表象中用于描述的词藻不是以点状形式定义的图象中的灰度，而是如形状、形态、运动和空间分布这样的描述。由感知组织完成的这样的变换可被看作与对实函数作 Fourier 变换相似。在作 Fourier 分析时，一个函数是以 Fourier 域中的 Fourier 分量来表示的。利用 Fourier 分析，我们可以用一组 Fourier 系数来描述一个函数。这样做的优点是用一组有限的系数就可提供一个良好的整体描述，这样使复杂性大为降低。虽然，很可能这个函数没有一个点的值是被正确地表示出来。这里就象是在感知中那样，局部与整体虽然是相互联系的，但本质上是不同的。总之，感知组织对传感器数据进行了整体的分析，得到一组宏观的表象。这样的宏观表象就是我们在进行认知活动时使用的基本构件，用它们可构成我们对外部世界的描述。

Gestalt 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。因此自从在本世纪二十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 80 年代以后，Witkin 和 Tenenbaum[Win 83]，Lowe[Lowe 86]，Pentland[Pen 86]等人在感知组织的原理，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。

### 1.2.3 现有视觉理论的革新

如前所述计算机视觉研究的发展开始于 60 年代初，在基本研究方面取得显著进展是在 70 年代末和 80 年代。这主要归功于 Marr 的视觉计算理论的推动。这个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、临床病理神经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。Marr 理论的出现无论对人工智能研究和神经科学的发展都产生了深远的影响。Marr 理论的出现使得 80 年代的计算机视觉的研究与以前相比有显著不同。主要表现在研究内容和方向集中在与人类视觉系统中的感知独立模块相对应的课题上，也就是根据影调、运动、立体、轮廓、纹理等线索恢复物体表面的形状。这些研究极大地深化了计算机视觉的研究。但是 Marr 的视觉计算理论还不能被认为是一个完善的理论。它没能反映人类视觉的某些重要的本质，这就是人类视觉中的选择性和整体性。

人类视觉最显著的特点之一是有选择性。这是指观察者的注意力总是有目的地指向他最感兴趣的事物。一般生物最注意的是环境中时常变化的事物，忽略固定不变的事物。因为这样就可以迅速辨别出什么是对自己有益的，什么是对自己有害的。从而作出攫取或躲避反应。另一个重要的特点，如 Gestalt 心理学家发现的那样，是人类具有对图象数据进行组织归纳的能力，也就是在多个层次上发现图象数据的规则性（regularity）、一致性（Coherence）、连续性（Continuity）等整体特性的能力。实验证明，人类视觉系统具有在低层处理中获取

图象拓扑特性的能力[钱学森 86]。

Marr 的理论完全不考虑视觉中的选择性和整体性，把初级视觉研究的目标确定为按照各种物理模型和附加约束条件，根据图象中各点灰度或其它测量结果，恢复景物中表面的有关特性，如表面方向、深度、反射率等。但由于图象中各点的灰度是光照，表面材料的反射特性、表面方向、观察方位等多种因素共同作用的结果。并且在成象过程中失去了各点的距离信息，所以，根据图象中的测量值（如灰度）恢复相应表面的三维特性（如，深度、方向），从本质上来说是一个约束不充分（underconstrained）的问题。也就是说，图象的测量值本身不能提供充分的信息来恢复相应表面的三维信息。因此，为能根据 Marr 理论恢复表面的三维信息必须增加附加的约束条件。例如，把物体仅限于刚体的范围，假设表面是连续的，各向同性的；或更为特殊的约束，如表面是由平面构成，点光源照明，材料的反射率为常数等。这些约束条件只能在某些人造环境下（例如在所谓的“积木世界”）得到满足，而在自然界或实际情况下通常是不满足的。而且即使具备了这些条件，目前采用的大多数求解方法类似于求解经典的边值问题。总的来说性能比较脆弱，容易出错。Marr 理论的这些困难在 80 年代末已经暴露得较为明显。

由上述分析可知，现有的两种视觉信息的处理理论各自遇到了严重的困难，还都不能自成系统地、可靠地处理视觉问题。因此，有的研究者提出了各种设想对上述理论提出了修改，并试图把这两种方法以取长补短的方式结合起来［witkin 83］。

一种改进的设想是基于模型的视觉理论［Gib 67］［Pen 86］［Gib82］。这种理论认为信息的概念是与从一组候选的对象中作出选择相联系。如果不知道一组可供选择的刺激或响应，人们就对刺激或响应无从说起。此外，人们还必须知道定义这组候选物和对这组候选物的成员进行区别的特性或特征。而且随着要解决的任务不同，这些特征和特性也不同的。例如，在视觉敏感中，刺激引起在视网膜的一组可能的状态中选择一种状态，并得到一幅图象。在感知中，选择是根据不变量（Constancies）和参数（Parameters）作出的。如果一个婴儿能听到声音，但他的感知不变量只包括“安静”和“噪声”的话，那么任何音乐对他来说者将包括一样多的信息，而这些音乐对一个训练有素的音乐家来说就会包含丰富得多的信息。

此外这种视觉理论利用特征检测器的概念作为把点状的图象数据与宏观信息相联系的桥梁［Pentland 86］。因此，基于模型的视觉理论体现了 Gestalt 理论中的选择和整体性。

另一种改进的设想是连接主义模型（Connectionist model of vision）［Fel 80, 82, 85］［Bal 84, 86］。动物的大脑进行计算的方式不同于当前传统的串行计算机。动物神经单元的计算相对是比较慢的。但它们之间具有复杂的并行连接，形成高度的并行计算结构。当前神经科学中的许多研究都是关于探索这些连接，以及试图发现这些连接是如何传递信息的。视觉的连接主义理论的基本前提是认为单个神经元并不传递大量的符号信息，而是通过与许多相似的神经元以适当方式相连接来完成计算。从点状的图象数据变换成一个整体的描述需要大量的计算，如前所述，这对目前的串行计算机来说是难以承受的。而上述并行计算结构则提供了一种可能的途径。连接主义模型的视觉理论认为 Hough 变换起重要作用。Hough 变换利用样板或模型（即圆周、直线、和其它几何形状）和参数（变量）来完成点状传感器数据到整体描述的聚集。 此外，Hough 变换从本质上来说是适合于由并行结构来实现。Ballard 还提出了连接主义模型的计算结构，详见［Bal 84］。

### 1.2.4 感觉的解析计算模型［Mar83］

目前数字计算机已能代替人完成复杂的科学计算， 其速度远超过人脑。并且现在已研制出能在比较窄的领域里表现出成年人推理能力的程序。但目前由计算机控制的智能机器在感觉能力方面表现出来的水平大致只能与蚱蜢相当。人们在这些领域里所作的努力是差不多

的，但结果却有这么大的差别，其原因是什么？此外，人类感觉的反应是极其迅速，并且非常可靠的。这可能意味着，感觉系统的工作更象是从某种相关的记忆里作回想，而不是进行了某种计算。这里所说的是指由模型化为“图灵机”（Turing Machine）或其等同物的计算。因此，这就很自然地产生一个问题，感觉过程能否模型化为图灵机，并用数字计算机来实现呢？也就是说，感觉过程的模型能否建立在符号计算的基础上。对人脑和神经生理的许多研究成果表明解析计算（Analytical Computation）能更好地反映感觉过程的某些重要特征。Hopfield 的神经网络模型就是一种解析计算模型［Hopfield 82］［Hopfield & Tank 85］。

这种模型认为人脑中进行的信息处理不是离散的符号处理，而是一种连续的反应过程。这就是说，神经网络中的神经元之间的信息交流在除了最后一层以外的所有层次上都是连续的。来自环境的输入信息引起网络的进化过程，这个过程是由某种最小值原则导引的。系统的稳定状态就是系统的局部最小值。这样的连续过程服从解析计算模型的微分方程。目前这方面的研究已引起广泛的兴趣，并已取得不少有意义的成果。

## 1.3 计算机视觉研究的发展

人类对数字图象的研究早在本世纪初的 20 年代、从 Bartlane 建立电缆图片传输系统，完成横跨大西洋的图片传输就开始了［che 68］。但对计算机视觉的研究则是在六十年代中期、从 Roberts［Rob 65］对积木世界的研究才开始。在这二十多年的发展过程中计算机视觉研究的内容、方法和理论都发生了很大变化。介绍这方面的发展过程，将有助于了解计算机视觉研究中的基本矛盾、发展趋势和解决途径。

### 1.3.1 自底向上的视觉处理方法

通常认为 Roberts［Rob 65］在 1965 年发表的论文是计算机视觉研究中的开拓性工作。他研究了根据图象来理解由多面体积木块构成的景物（以后常称为积木世界）的方法。为此首先要把输入图象（图 1.2(a)）转换成线画图（图 1.2(b)）。具体过程是：预处理去除噪声；对图象灰度作一阶空间微分，选择灰度微分值高的象素作为边缘点；连接相邻边缘点，并对短的边缘作平滑处理（图 1.2(e), (f)）；然后把边缘点用最小均方差的方法连接成直线。下一步是根据矩形块和三角块这样的基本多面体（称为基元）来解释所得到的线画图。景物中的一个简单物体可看成是这些基元经过变换以后得到的一个实例。变换包括沿三个轴的旋转、比例和投影等。组合的多面体可看成是由若干个简单的多面体粘合而成。识别线画图是由哪些基元构成的方法是把线画图与基元的拓扑特性（面、线、顶点的结构）进行

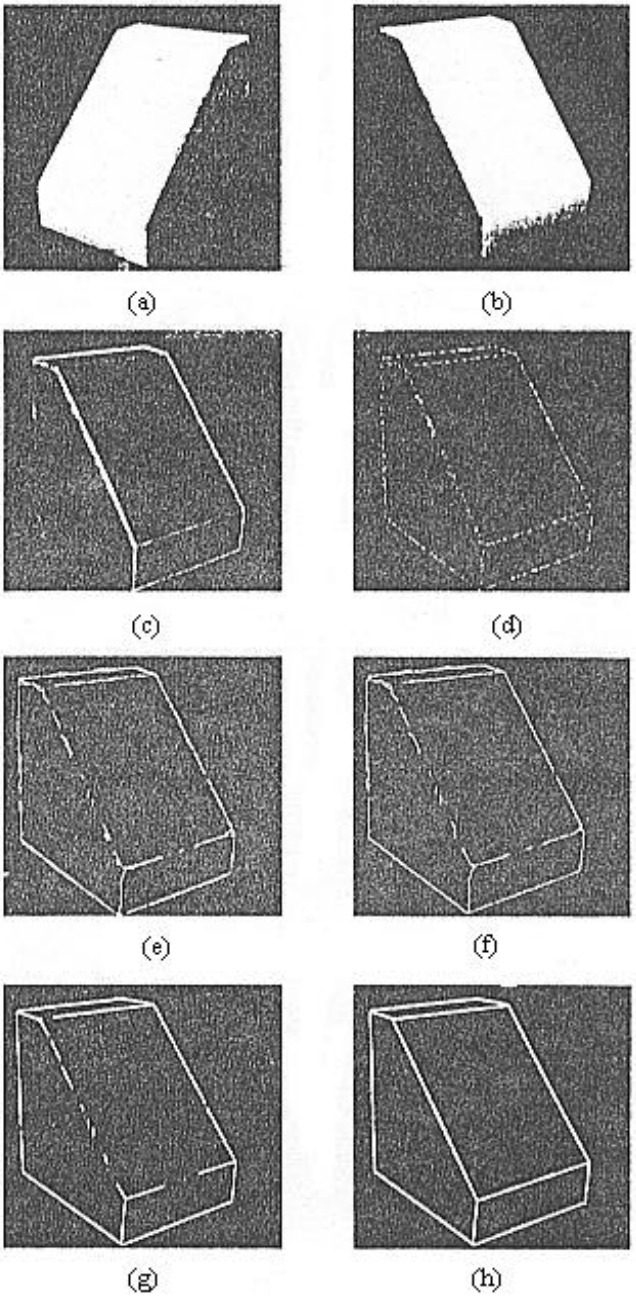


图1.2　Roerts的线画图抽取过程

匹配，先是试探性的匹配，然后由一个量度过程来确定这个变换是否合理。当在景物中识别出变换的基元时，这个基元就被想象成已被切开，并被移走了。新的可看见的线条就填进来。然后对这个余下的线画图作进一步的分析直到完成全部的分析。图 1.3（$A_1$）中所示为原始的景物线画图，识别的方法是试图用矩形块和三角块的模型与此线画图相匹配（按惯例先试矩形块，然后再试三角块）。把矩形块模型投影到线画图上（见图 1.3($A_2$)，并把转换的模型放入三维结构（见图 1.3($A_3$)）。模型识别以后，用过的点和线就被移走，这就产生了 $B_1$。接着就进行新的搜索，寻找相匹配的模型。这次发现底部的矩形块，这就产生了 $B_2$ 和 $B_3$。下一次是矩形块模型与 $C_1$ 中的左面相匹配，于是产生 $C_2$ 和 $C_3$。最后在 $D_1$ 中剩下一个三角块，当把三角块模型加到三维结构中去时，就产生了 $D_3$ 中所示的完整的三维模型，根据 $D_3$ 就可以生成任意的视图。在 Roberts 的程序中已经包括了目前视觉程序中的大部分组成：预处理、边缘检测、线条拟合和建立线画图、物体建模和匹配。他提出的识别过程是按从低层到高层、

从图象到物体的顺序进行的。在早期的计算机视觉研究中许多都是以积木世界作为研究对象，并且大多采用这种自底向上（bottom-up）的方法。

## 1.3.2 图象分割

上述自底向上的方法能否成功主要取决于是否能成功地完成低层和中层的处理。因此，在计算机视觉研究的初期发展了许多从图象中抽取边缘和线条的技术。这些技术大多与Roberts的相似。通过在各个象素周围窗口中进行的局部运算可确定这个象素是否在边缘上。通常所作的运算是微分或与理想的边缘作相关运算。然后对所的结果再作阈值运算，以产生二值图象。对各种边缘检测算子的介绍可参阅［Ros 82］。被检测到的边缘点要连接成有意义的线段。这可通过按某种规则对边缘点进行跟踪来完成。在跟踪过程中会遇到边缘之间的间隙，曲线的突然变化或由于噪声引起的假边缘点等情况。这将要求复杂的跟踪规则。

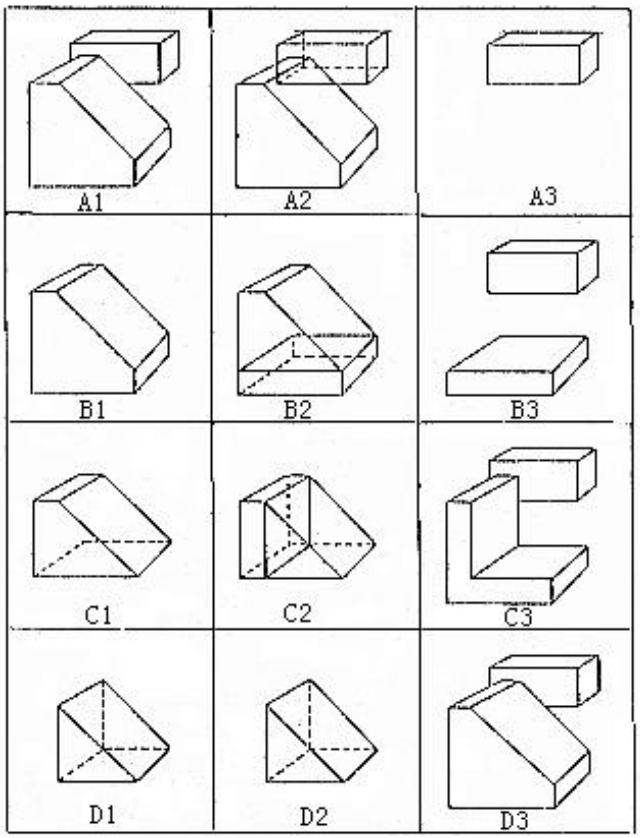


图1-3　复合物体三维模型的构成，景物的原始线画图$A_1$，经过步骤A, B, C和D中4个模型的依次识别、删除，得到$D_3$中所示三维结构。

把图象分割成有明确含义的（代表各种物体或背景）区域的方法除了上述检测两个区域之间边缘的边缘检测方法以外，另一种是区域分割法。它通过寻找其中的象素具有相同的灰度（或特性）的区域来完成图象的区域分割。这两种方法是互补的，在理想情况下得到的结果是等价的。

图象分割的目的是按有明确含义的线段或区域来描述图象，以便与模型相比较。在计算机视觉的早期研究中对分割方法作了大量研究。结果表明分割是非常困难的。例如，图象经过边缘检测算子运算以后要作阈值运算以判断边缘点。在选择阈值时，如阈值取得低则将能检测出灰度对比度低的边缘点，但也会把由于模糊或表面不均匀造成的虚假边缘检测出来。相反如把阈值取得高些，则会在抑制噪声的同时也漏失真正的边缘。在作区域分割时也会因阈值不合适把同一物体表面分成几个区域或把不同的物体表面当作一个区域。通过对环境条件作较为严格的限制和仔细调整阈值参数可缓解这些困难，但是当遇到同一区域中的灰度（或特性）不同，而具有相近灰度的象素又不属于同一区域的情况，只是靠调整阈值就难以解决问题。例如，即使在均匀光照的情况下，一个曲面（如鸡蛋表面）上各点的灰度也不是均匀的，而是呈现灰度影调的分布，这时就难以选择阈值。这时为使分割能得到有明确含义的区域就需要具有外部的自顶向下（top-down）的知识，和可能还需要在实践中总结的经验性知识，即所谓的启发式知识。

## 1.3.3 利用启发式知识的方法

在图象理解中首先广泛利用启发式知识的是 Guzman［Guzman 68］。他建立的“See”程序能把线画图分割成三维的物体。他的方法是对线画图中的由交汇的直线构成的顶点类型作分类。他发现根据顶点的类型就可以得到关于物体区域之间关系的局部线索。例如，Psi 类型的顶点（图 1.4，指交汇在此顶点的几条线形成Ψ形状）经常出现在两块柱状积木对准摆放时。这意味着上半部的两个区域属于一个物体，下半部的两个区域属于另一个物体。Guzman 把属于同一物体的两个区域之间的连接称为链，并用链来表示关于区域之间连接关系的启发式规则。See 程序根据这些关于顶点的启发式知识来连接各个区域。通过许多链连在一起的各个区域可能同属于一个物体，而属于不同物体的区域之间没有链或只有极少数的链连接。但这样的用于连接的启发式知识还不能最终确定哪个区域属于哪个物体。所以 Guzman 设计了根据区域之间链的数量、强度和拓扑关系来分配区域的启发式规则。他的算法可分析一幅相当复杂的线画图。这是早期视觉处理中应用启发式知识所取得的显著成就之一。它证明可以通过符号处理，而不是通过匹配过程来解释线画图。但是 Guzman 的方法还有根本性的困难。虽然 See 程序可识别三维物体，但所用的启发式知识仅限于二维的图象领域；其次，这些启发式知识是非常专门的，不通用并缺乏物理基础。这两个问题不是 Guzman 的方法所特有的，而是这类利用启发式知识方法的通病。

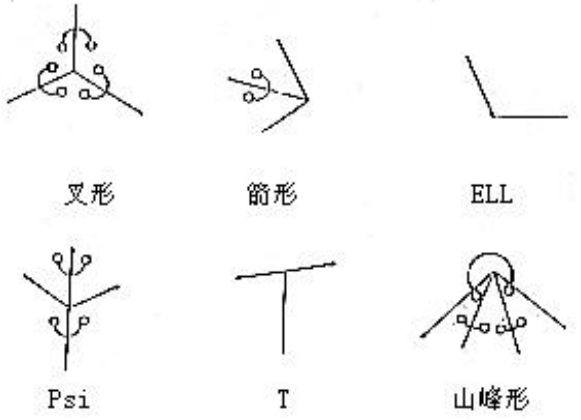


图 1.4　链围绕顶点的排列 AI & A　P.230, fig. 8-31

## 1.3.4 利用高层知识的方法

顺序的自底向上的视觉处理方法遇到的严重困难是图象分割。甚至在最简单的积木世界的情况下要抽取完美无缺的线画图也几乎是不可能的。事实上，存在一个大家都熟知的公理，只有理解了线画图以后才能得到完美的线画图。解开这个死循环的方法之一是利用关于特定物体或特定种类物体的专门知识来帮助解释输入数据。这些专门知识也可称为是语义知识，并一般以物体模型的形式来表示。这里需要着重指出的是在视觉处理中应用专门的语义知识应与应用关于物体的通用物理特性，如连续性知识相区分。一个视觉处理方法如果应用关于物体的通用物理特性的知识，那么还被认为是自底向上的方法。

Fal［Fal 72］的 INTERPRET 程序利用模型来帮助对不完全线画图进行解释。它先分析一幅线画图，并对图中的物体种类和方位作出假设。然后预测所假设景物的线画图，并最后试图校验所假设的线画图。这种主动地应用模型的方法被称为基于模型的、自顶向下的、语义的、和目标驱动的方法。在某些系统中模型只是简单地校验结果，而在其它一些系统中则完全由模型来控制在何处可见到何物。在 70 年代的初期，学术界普遍认为不引入自顶向下的景物知识的初级视觉处理从本质上看无法产生丰富的、有用的描述。并认为一个聪明的视

觉系统应该对景物中有些什么事物了解很多。这类方法中有代表性的程序有 Shirai 的语义边缘检测方法［Shir 73］、Yakimosky 的基于语义的区域分析方法［Yak 73］，和 Tenenbaum 的解释导引的分割方法［Tene 76］等。但这些方法在高级语义知识如何与低级的聚集处理相衔接上遇到严重的困难，并且不可避免地要引入一些过于简单的假设，因此并未取得预期的成功。这样，在七十年代后期计算机视觉研究面临左右为难的困难境地。一方面认为没有关于图象是什么物体的先验知识，就难以达到完善的图象分割，也就难以理解图象；另一方面在应用先验知识来导引低层处理进行图象分割时又遇到了上述的严重困难，于是很自然地提出一个问题：在不知道图象中有些什么物体的条件下，二维的灰度图象到底能提供什么样的信息。例如，人是通过用双目观察事物来获取深度信息的，那么人是在识别出看到的是什么物体以前就获取了深度信息呢？还是不论是否知道看到的是什么物体都能获得深度信息？再例如，灰度图象能否提供物体的形状信息等。对这些问题的回答，不可避免的地要涉及对人类视觉的研究。

## 1.4 人类视觉与计算机视觉的比较

目前人们所建立的各种视觉系统极大多数是只适用于某一特定环境或应用场合的专用系统，而要建立一个可与人类的视觉系统相比拟的通用视觉系统是非常困难的。主要原因有以下几点：

1. 图象对景物的约束不充分。首先是图象本身不能提供足够的信息来恢复景物，其次是当把三维景物投影成二维图象时丧失了深度信息。因此，需要附加的约束才能解决从图象恢复景物时的多义性。

2. 多种因素在图象中相互混淆。物体的外表受材料的性质、空气条件、光源角度、背景光照、摄象机角度和特性等因素的影响。所有这些因素都归结到一个单一的测量，即象素的灰度。要确定各种因素对象素灰度的作用大小是很困难的。

3. 理解自然景物要求大量知识。例如，要用到阴影、纹理、立体视觉、物体大小的知识；关于物体的专门知识或通用知识，可能还有关于物体间关系的知识等。由于所需的知识量极大，难以简单地用人工进行输入，可能要求通过自动知识获取方法来建立。

4. 人类虽然自己就是视觉的专家，但它又不同于人的问题求解过程，难以说出自己是如何看见事物，从而给计算机视觉的研究提供直接的指导。

视觉机理的复杂深奥使有些学者不禁感叹道：如果不是因为有人的视觉系统作为通用视觉系统的实例存在的话，他都怀疑不能找到建立通用视觉系统的途径。从另一方面来看，正如 Rossen 所说的：“从进化的观点来说，生理系统是人类解决复杂问题的最好的百科全书。”对建立通用视觉系统这个难题来说，在长期进化过程中高度发展了的人类视觉系统确实提供了最好的实例。例如，对人类视觉中可区分的独立视觉模块的研究就帮助我们认识了在没有先验知识的条件下初级视觉处理能否获得丰富的有用描述的问题。

许多心理物理学实验证明在人的视觉系统中似乎存在独立的特定模块。其中著名的例子有 Land[Lan 71]关于照明的计算和 Julesz[Jul 71]关于立体视觉的实验。例如，Julesz 的实验证明了人的视觉系统可在对图象内容一无所知的情况下进行立体视觉处理，获得深度信息。他用计算机产生左、右两幅由随机点组成的图象。从单幅图象看，这两幅图都不过是由一些随机分布的点组成的图案，并无物理含义。但当用立体镜观察，把这两幅图融合在一起时就得到了立体信息（详见第五章）。这说明在人的视觉系统中似乎存在着独立的立体视觉模块。除此以外还有其它的独立模块。Horn[Hor 75][Ike 81]研究了人类视觉根据影调得到物体形状的能力，steven[ste 81]研究了人对表面轮廓作出三维形状解释的能力等。更多的有关研究可参见[Bob 81]。显然，这些研究结果与 70 年代初期流行的认为初级视觉处理难以得到丰富的有用描述的观点

相反，并且标志着 80 年代的计算机视觉研究的趋势与 10 年前已大不相同。其主要特点是研究集中在与人类视觉系统中可区分的独立模块相应的课题上，并且许多研究者希望他们的工作能与心理物理学和神经生理学的理论有直接的联系。从长远来看，建立人类视觉的计算理论，并进而建成可与人类视觉系统相比拟的通用视觉系统是计算机视觉研究的最终目标。对人类视觉的研究涉及神经生理学、心理物理学、心理学等多方面。对人类视觉机理的了解为建立视觉的计算理论提供有益的启示。与此同时，视觉计算理论的研究又促进了在上述领域中引入计算机技术，这又推动了这些学科自身的发展。

强调计算机视觉研究与人类视觉研究之间的紧密关系，并不意味着计算机视觉系统要机械地模仿人类视觉系统。因为生物视觉系统是生存竞争中进化的产物，带有由此而来的优点和局限性[Bra 83][Per 83]。影响生物或人类视觉系统的因素有：

1. 能否根据距离的远近对不同的物体，特别是对不同的生物作出不同反应的能力应该是视觉系统发展中的重要准则。视觉系统的优点在于不与环境直接接触就可以作出响应。如果视觉系统不能反应距离远近，那么就使视觉系统的这个优点受到严重限制。

2. 应用视觉的初等生物体需要有能力自动地对环境的刺激作出响应。理解能力是进化的更高阶段中感知过程的结果。

3. 感觉机制必然是在足以确保生存或对生存有利的基础上发展起来的。因此不能认为这些机制是利用了从数学或计算机观点来看的最优解。所以有必要具体分析人类视觉系统的特点。人类视觉系统大致有以下特点：具有高分辨率，有立体观察、优越的识别能力和灵活的推理能力，可灵活地根据各种视觉线索进行推理，即

(1) 深度感觉的首要性。可能初等生物体表面上的光敏区开始时只能提供关于光源的方向信息，或者阴影可能表示一个捕食动物正在逼近的方向。感光区域上感知的阴影面积的增加可能意味着捕食者正在逼近，这可能是深度感知的早期形式。这样的进化过程只是一种分析，但是有依据说明在人类视觉系统中探测阴影逼近的机能直接与感知物体在深度上的移动有关。对人类试验者进化的心理物理学实验和对猫进行的生理学实验都支持这种机理的存在。这种类型的机理与 Marr 提出的从视网膜上的二维表象开始，经过一个或多个中间表象计算再作三维解释的机理不同。

(2) 感知是个自动进行的过程。感知的特点在于它是一个自动进行的过程，并且它抵制根据与其相矛盾的知识作出修改。实验证明，如果让一个观察者先观察一个旋转着的收缩螺丝，那么在他习惯以后再去观察另一个物体，例如，一张人脸，就会感到人脸在膨胀。观察者可能已经知道人脸并没有膨胀，但这并不妨碍得到这种膨胀的感觉。还有许多例子可证明人会出现这种明知与常识或已知情况相矛盾的感觉。

虽然与感觉相矛盾的知识不能改变人的感觉，但显然它可影响人对视觉刺激作出不同的反应。一个人如果根据情况已知不会有大的物体正在逼近，那么当他看到出现一个影子时不会逃跑。但当影子突然出现时，他还会不自觉地感到害怕。人类虽然已具有较高级的理解能力，但视觉系统似乎保留着对某种刺激自动作出反应的能力。从进化的观点，感知与知识相分离可能是有道理的，但对用于准确分析三维景物的视觉系统来说就是不可取的。

(3) 感知中对启发式知识的应用。自然环境中的许多物体是刚体，所以在进化过程中发展起来的人类视觉系统在根据视网膜上的物体投影分析物体时假设物体是刚体，以简化分析是有道理的。例如，如果视网膜上成象的大小变化，而形状保持不变，就可认为物体的远近起了变化。但在某些特殊情况下，在分析成象的大小变化时视觉系统并不采用通常的刚性物体的假设，而是采用不同于刚性假设的其它特殊假设。例如，当在平面中旋转图 1.5(a)所示的由两个螺纹状图〔图 1.5(b)和(c)〕连接成的图形时，人在观察时通常会感到这是一个正在变形的三维形状。这种情况下在视网膜上的成象是与刚体的运动不相符合的，也即图形在平面内旋转，而这个运动的刺激图象似乎给人以图形的一部分正在膨胀，而另一部分正收缩的

印象。在刚体上是无法同时造成收缩和膨胀印象的。因此，这给人的感觉是一个正在变形的三维物体。这种感觉并不因为刚性物体的假设而消失。

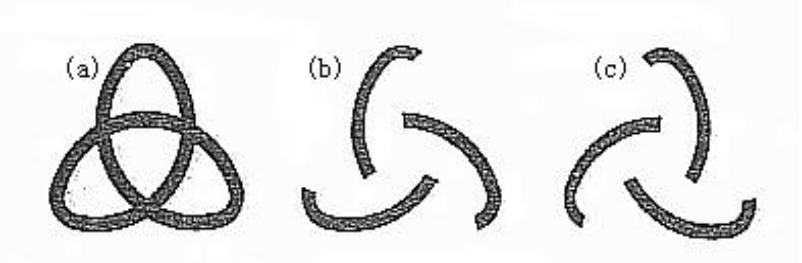


图 1.5 (a)(b)(c)这个二维图形在平面内旋转时，使人感到是一个非刚性的三维形状，(b)、(c)，这两个图象在平面内旋转时似乎在膨胀或收缩，这要取决于旋转的方向。

在计算机视觉系统中如果在计算中保持物体的刚性假设就可以避免上述这种不可靠的感觉。虽然人类视觉中的这种近似有时会造成错误，但比单纯应用刚性假设可能有以下优点：(1) 这样的近似使系统不致于过分偏向于刚体运动的感知，不然就会反过来影响对非刚体的感知；(2) 人类视觉系统所应用的启发式感知的过程对生理系统来说可能要比采用通用的刚性物体假设简便；(3)人类视觉系统对几何问题不采用严格求解方法的原因是这样可以更为迅速地求解。对动物的生存来说，一个可迅速地探测到潜在危险的近似求解方法比虽然可准确识别，但费时过多的方法要有用得多。

除了以上特点以外，人类视觉系统有分辨率高、识别能力优越、能进行灵活的推理等优点，以及在根据外部成象推论物体三维形状或姿态时会产生严重错误的弱点[Perkins 83]。综上所述，从建立通用的计算机视觉系统的角度来看，关键点不是机械地模仿人类视觉系统，而是通过对人类视觉系统的研究发现是什么因素使人类视觉系统的性能如此之好，并且把它结合到计算机视觉系统中去。

参考书：

[钱学森 86]钱学森，关于思维科学，上海人民出版社，1986。

[Arn 69] Arnheim, R., Visual Thinking, University of California Press, 1969.

[Bal 84] Ballard, D.H., Parameter Nets, Artificial Intelligence 22(1984), 235-267.

[Bal 86] Ballard, D.H., Cortical Connections and Parallel Processing: Structure and Function, Behavioral and Brain Sciences 9, 1986, 67-120.

[Bra 83] Braunstein, M.L., Contrasts between human and machine Vision: Should technology recapitulate phylogeny, in Human and Machine Vision, Bech, J. et.al.eds, Academic Press, 1983, 85-95.

[Fal 72] Falk, G., Interpretation of Inperfact Line Data as a Three-dimensional Scan, Artificial Intellligence 3, 1972, 101-144.

[Gib 67] Gibson, E, J., Principles of Perceptual Learning and Development, New York, 1967.

[Gib 82], Gibson, J.J., What is Involved is Surface Perception, In J.Beck(ed).Organization and Representation in perception, 1982.

[Guz 68] Guzman, A., Decomposition of a Visual Scene into Three-Dimensional Bodies, in Proceedings of AFIPS Fall Jaint Conference, 33: 291-304.

[Hop 82] Hopfield, J.J., Neural Networks and Physical Shstems with Emergent Collective Computational Ability, In Proceedings of the National Academy of the USA , 79, 1982, 2554-58.

[Hop 85] Hopfield, J.J., and Tank, D.W., Neural Computation in optimization Problems, Biological Cybernetics, 52, 1982, 141-152.

[Hor 75] Horn, B.K.P., Obtaining Shape from Shading Information, In The Psychology of Computer Vision, P.H. Winston(ed), McGraw Teill Book Co., New York, 1975, 115-155.
[Low 85] Lowe, D.G., Perceptual Organization and Visual Recognition, Kluwer Academic Publishers, 1985.
[Ike 81] ikewchi, K. & Horm, B.K.P., Numerical Shape from Shading and Occluding Boundaries, Artificial Intelligence, 17. 1981, 141-184.
[Jul 71]Julesz, B., Foundations of Cyclopean Perception, Chicago University Press, 1971.
[Kan 87] Kanal, L. & Tsao, T. Artificial Intelligence and Neural Perception, In Intelligent Auto nomous Systems, 1987.
[Lef 79] Lefton, L. A., Psychology, Allyn and Bacon. Inc. 1979.
[Mar 82] Marr, D., Vision, W.H. Freeman and Company, 1982.
[Neg 91] Negahdaripour,S. & Jain, A.K., Final Report of the NSF Workshop on the Challenges in Computer Vision Research; Future Directions of Research, Lahina, Maui, Hawaii, Jane 78, 1991.
[Per 83] PerKins, D.N., Why the human perceiver is a bad machine, in Human and Machine Vision, Beck, J.et.al.eds, Academic Press. 1983. 341-364.
[Pen 86] Pentland, A.P., Part Models, In Proceedings of Int. Conf. Pattern Recognition and Computer Vision, MiamiBeach, Florida, June 22-26, 1986, 242-249.
[Pen 88] Pentland, A.P., The Parts of Perception, in Advance in Computer Vision, vol.II, eds. by Brown, C., Lawrence Erlbraum Associates, 1988.
[Rob 65] Roberts, L.G., Machine Perception of Three-Dimensional Solids, in Optical and Electro-Optical M.I.T.Press. 1965, 159-197.
[Ste 81] Stevens, K.A., the information Contents of texture gradients. Biological Cybernetics, 42, 1981, 95-105.