第5章

计算机视觉

5.1 计算机视觉的基本概念与原理

5.1.1 计算机视觉的定义与研究目标

计算机视觉(Computer Vision,CV)是人工智能领域中一门以“让机器拥有视觉能力”为核心的前沿学科。简单来说,它试图通过技术手段让计算机像人类一样“看懂”世界:具体而言,是利用摄像头、传感器等设备替代人眼采集图像或视频数据,再通过算法让计算机对这些数据进行分析、处理,最终完成对目标的识别、跟踪、测量等任务,甚至将处理后的信息转换为更便于人类观察或仪器检测的形式。例如,在工业质检中,计算机视觉系统可以通过摄像头拍摄产品表面图像,自动识别肉眼难以察觉的微小划痕,这便是“机器视觉”在实际场景中的典型应用。

从学科属性来看,计算机视觉是一门高度交叉的综合学科。它以计算机科学为基础框架(如算法设计、数据处理),以数学为核心工具(如图像变换的矩阵运算、特征提取的概率模型),以物理学为底层支撑(如光学成像原理、色彩形成机制),同时借鉴了神经科学中人类视觉系统的工作机制(如视觉信号的层级传递与处理)。这种多学科的融合特性,让计算机视觉既需要工程化的技术实现能力,也依赖基础科学的理论突破。

计算机视觉的核心定义可以概括为:通过计算机对图像或视频数据进行采集、处理、分析和理解的完整流程,模拟人类视觉系统对客观世界的感知与认知能力。这里的“感知”指获取物体的基本属性(如形状、颜色、位置),“认知”则是在此基础上形成更高层次的理解(如判断物体的用途、预测物体的运动趋势)。

计算机视觉的研究目标并非单一维度的技术实现,而是包含从基础到高级的三层递进式目标,每一层都为上层提供支撑,共同构成完整的视觉能力体系。

1. 基础层面

在基础层面,研究重点是对图像的像素级信息进行处理,目标是优化原始图像质量,为后续分析扫清障碍。原始图像往往因采集设备(如摄像头分辨率不足)、环境干扰(如光线过暗、雨雪遮挡)或传输过程(如信号噪声)存在缺陷,因此需要通过算法进行修正。例如,在安防监控场景中,夜间摄像头拍摄的图像可能存在严重噪点,基础层的“降噪算法”能过滤这些干扰;在遥感卫星图像中,“图像增强技术”可以提升模糊区域的清晰度,让地面建筑的轮廓更易识别。因此这一

层的核心是“改善数据质量”，相当于为机器视觉“擦亮眼睛”。

2. 中级层面

中级层面的目标是在高质量图像的基础上，完成对图像内容的结构化解析，具体包括物体检测、图像分割和特征提取。物体检测旨在定位图像中感兴趣的目标（如识别画面中所有的行人）并标记其位置；图像分割则更进一步，将图像按语义划分为不同区域（如将“道路”“车辆”“天空”分割为独立区域）；特征提取则是从目标中提取关键信息（如人脸的五官轮廓、汽车的车牌特征）。这一层的核心是“理解图像内容的组成”，相当于让机器能“认出画面中的元素”。例如，在手机拍照的“人像模式”中，系统先通过检测定位人物，再通过分割区分“人物”与“背景”，最终实现背景虚化效果，这正是中级层面技术的典型应用。

3. 高级层面

高级层面是计算机视觉的终极目标之一，旨在实现对视觉信息的深度理解与决策推理。它需要结合中级层面的结构化数据（如物体位置、特征），结合场景知识进行综合分析，最终形成对场景的整体认知、行为预测或决策判断。例如，在智能家居中，摄像头识别到老人摔倒（中级层）后，系统能判断“需要紧急救助”并自动拨打求救电话（高级层）；在无人机巡检中，识别到输电线路的异常发热点（中级层）后，系统能进一步分析“可能存在线路老化风险”并生成维修建议（高级层）。这一层的核心是“基于视觉信息的智能决策”，相当于让机器能“思考画面背后的意义”。

4. 典型案例

以自动驾驶场景为例，这三层目标的协同作用尤为明显：基础层面通过“去雾算法”处理雨天摄像头拍摄的模糊图像，确保画面清晰；中级层面通过目标检测识别出前方的车辆、行人、交通信号灯，并通过分割技术区分车道线与路沿；高级层面则结合这些信息，判断“前方行人正在横穿马路”，进而决策“立即减速并开启警示灯”。从像素处理到行为决策，三层目标层层递进，共同支撑起自动驾驶的视觉感知系统。

5.1.2　计算机视觉与人类视觉的异同

人类视觉与计算机视觉，本质上都是“通过接收光信号理解世界”的信息处理系统——前者是经过亿万年进化形成的生物系统，后者是人类用技术构建的人工系统。二者在目标上高度一致：都需要将光信号转换为有意义的认知（如“这是一只猫”“前方有障碍物”），但在实现机制、优势局限上存在显著差异，同时也存在奇妙的相似性。

1. 核心共性：层级化的信息处理逻辑

无论是人类视觉还是计算机视觉，对信息的处理都遵循“从原始信号到抽象认知”的层级递进逻辑，这种共性也是深度学习模仿生物视觉的核心依据。

人类视觉的层级处理机制早已被神经科学证实：当光线进入瞳孔，首先由视网膜的感光细胞（视杆细胞、视锥细胞）将光信号转换为神经电信号（这一过程相当于“原始数据采集”）；随后信号传递到大脑初级视觉皮层，这里的神经细胞专门识别边缘、线条、明暗等基础特征（如“检测到一条水平边缘”）；接着信号向更高层级的视觉区域传递，基础特征被组合成更复杂的形态（如“边缘构成了圆形轮廓”）；最终在大脑联合皮层完成抽象认知（如“这个圆形轮廓是苹果”）。整个过程就像“搭积木”，从最小单元逐步构建出完整认知。

现代计算机视觉（尤其是基于深度学习的方法）刻意模仿了这种层级结构。以识别图像

中的“汽车”为例：神经网络的底层卷积层负责检测像素级的基础特征（如颜色梯度、简单纹理）；中层网络将这些特征组合成部件（如“车轮的圆形轮廓”“车窗的矩形边缘”）；高层网络则整合部件信息，最终判断“这些特征组合符合汽车的定义”。这种“底层特征→中层部件→高层类别”的处理流程，与人类视觉的神经信号传递逻辑高度吻合。

2. 关键差异：从生物本能到技术实现

尽管目标与处理逻辑相似，但人类视觉与计算机视觉的“硬件基础”“能力边界”“认知模式”存在本质区别，这些差异也决定了二者在不同场景中的优势与局限。

1）硬件与信号传递：生物器官 vs 电子设备

人类视觉系统是一套高度集成的生物装置：瞳孔通过收缩调节进光量，相当于“天然的光圈”；视网膜上的 1.2 亿个视杆细胞（感知明暗）和 600 万个视锥细胞（感知色彩），能在极暗环境（如星光下）或强光环境（如正午阳光）中自适应捕捉信号；神经信号通过视神经以“生物电脉冲”的形式传输，速度虽不及电子信号，却能通过神经递质实现信号的灵活调节。

计算机视觉则依赖人工硬件：摄像头的 CMOS/CCD 传感器是“电子视网膜”，其像素数量（如 4800 万像素）决定了原始信号的精细度，但无法像人类瞳孔那样自适应调节动态范围——在逆光场景中，这类设备往往会出现“高光过曝、暗部丢失”的问题；原始数据通过数据线（如 USB、HDMI）以二进制形式传输，速度可达每秒数十吉字节（GB），但无法像神经信号那样自主修正误差；最终的处理依赖 CPU/GPU 的计算单元，其优势是“并行计算能力”（如 GPU 可同时处理数百万个像素），但必须遵循预设的算法逻辑，无法像大脑那样“灵活容错”。

如表 5-1 所示，二者的硬件对应关系清晰体现了“生物本能”与“技术模拟”的差异。

表 5-1　人类视觉和计算机视觉的作用和核心差异

人类视觉	计算机视觉	作　用	核心差异
瞳孔→感光细胞	摄像头→像素阵列	接收光信号	人类可通过瞳孔收缩、感光细胞敏感度调节适应光线变化；计算机依赖传感器固定参数，需算法补偿光线不足
视神经	数据线	传输原始数据	人类神经信号可通过生物机制修正干扰信号；计算机数据传输依赖物理线路，易受噪声影响（需额外编码校验）
大脑皮层	CPU/GPU	处理并提取意义	人类大脑可结合经验灵活解读模糊信号；计算机需严格遵循算法逻辑，依赖数据训练

2）信息处理能力：适应性 vs 精确性

人类视觉的核心优势是“天然的适应性与常识推理能力”。经过长期进化，人类能在复杂环境中自动聚焦关键信息：在喧闹的商场里，你能瞬间从人群中认出朋友的脸（忽略无关背景）；看到“一只猫坐在椅子上”的模糊图像，即使猫的部分身体被遮挡，也能结合“猫有尾巴、椅子有四条腿”的常识补全信息。这种“基于经验的模糊推理”，让人类视觉在光线变化、物体遮挡、姿态改变等场景中依然稳健。

计算机视觉则在“精确性与效率”上占据优势。它能在 1s 内完成对 10 万张图像的检索（如从监控录像中快速定位嫌疑人），这种“海量数据处理能力”是人类视觉无法企及的；在精度要求极高的场景中（如芯片质检），它能识别出 0.1μm 的划痕（相当于头发直径的 1/500），

而人类肉眼的极限分辨率仅为 50μm。

但计算机视觉的短板也十分明显：它缺乏人类的“常识储备”，在未经过训练的场景中极易出错。例如，给计算机输入一张“猫穿着衣服站立”的图像，如果训练数据中没有类似样本，可能会误判为“其他动物”；而人类即使从未见过这种场景，也能通过“猫的脸型、毛发特征”结合常识判断“这是一只穿衣服的猫”。此外，人类视觉对图像的理解是“整体性认知”（如看到“杯子”会同时联想到“可用来喝水”），而计算机视觉往往是“局部特征匹配”（如通过“圆柱形轮廓＋开口设计”识别为杯子，却无法理解其功能）。

3）学习模式：经验积累 vs 数据训练

人类视觉的“学习”是潜移默化的经验积累。婴儿从出生起，通过观察父母的脸、接触周围的物体，逐渐建立“形状→名称→功能”的认知关联，这个过程无须“海量标注数据”，却能形成普适性的认知（如见过一只狗后，能识别不同品种的狗）。

计算机视觉的“学习”则依赖大规模标注数据。要让系统识别“狗”，需要给它输入数万张不同品种、不同姿态、不同环境下的狗的图像，并人工标注“这是狗”，才能让算法记住“狗的特征”。一旦遇到训练数据中没有的情况（如从未见过的稀有犬种），识别准确率就会大幅下降。这种“数据依赖”是当前计算机视觉的核心局限——它无法像人类那样通过“举一反三”形成抽象认知。

3. 总结：互补而非替代

人类视觉与计算机视觉并非“谁替代谁”的关系，而是在各自擅长的领域形成互补。在需要“常识推理、灵活适应”的场景（如家庭教育中理解孩子的表情与情绪），人类视觉的作用无可替代；在需要“高效计算、精确检测”的场景（如疫情期间的人脸识别测温），计算机视觉成为核心工具。

随着技术发展，计算机视觉正不断借鉴人类视觉的优势——例如，通过“注意力机制”模拟人类“聚焦关键信息”的能力，通过“迁移学习”减少对标注数据的依赖。而人类也在通过计算机视觉延伸感知边界——如借助显微镜图像分析细胞病变（超越肉眼分辨率）、通过卫星图像监测森林火灾（覆盖广阔区域）。这种“生物智慧”与“技术能力”的融合，正是计算机视觉发展的核心动力。

5.1.3　计算机视觉的技术体系架构

计算机视觉的技术体系架构是一套从“感知物理世界”到“理解语义信息”的完整技术链条，它像一条精密的流水线，将原始的光学信号逐步转换为有价值的认知结果。这一架构可清晰划分为数据输入层、预处理层、特征提取层、高层理解层 4 个核心环节，每个环节既独立承担特定任务，又与其他环节紧密衔接——前一层为后一层提供输入，后一层以前一层的输出为基础，共同支撑起计算机视觉的完整功能。

1. 数据输入层：从光学信号到数字信号的“转换器”

数据输入层是计算机视觉系统与物理世界的“接口”，其核心任务是通过图像传感器采集光学信号，并将其转换为计算机可处理的数字信号（即图像或视频数据）。这一过程相当于为系统“采集原始素材”，直接决定了后续所有处理的基础质量。

常见的图像传感器包括两类：一类是用于实时采集的设备（如摄像头、工业相机），另一类是用于静态采集的设备（如扫描仪、卫星遥感传感器）。不同场景对传感器的要求差异极

大：在手机拍照场景中，摄像头需要兼顾便携性与色彩还原能力，通常采用小型 CMOS 传感器；在工业质检场景中，工业相机需具备极高的帧率（如每秒 1000 帧）和分辨率（如 1200 万像素），以捕捉高速运动零件的细节；在卫星遥感场景中，传感器则需能接收可见光、红外线等多种波段的信号，从而识别云层、植被等肉眼不可见的信息。

从技术原理来看，数据输入的过程可分为三步：首先，传感器中的感光元件（如 CMOS 中的像素单元）接收物体反射的光线，将光信号转换为电信号；随后，通过模数转换器（ADC）将模拟电信号转换为数字信号（即由 0 和 1 组成的像素值）；最后，将像素值按行列排列，形成二维的图像数据（如 1920×1080 分辨率的图像，包含约 200 万个像素点）。对于视频数据，还需通过连续采集（如每秒 30 帧）并记录帧间时序关系，形成动态序列。

需要注意的是，数据输入层并非简单的“采集”，还需根据场景需求进行初步适配。例如，在自动驾驶中，摄像头需配合激光雷达、毫米波雷达形成“多传感器融合”输入——摄像头负责识别交通信号灯的颜色，激光雷达提供三维距离信息，二者结合可弥补单一传感器的缺陷（如摄像头在暴雨天易受干扰，而激光雷达受天气影响较小）。这种“多模态输入”已成为复杂场景的主流方案。

2. 预处理层：为数据“提质增效”的“优化器”

原始采集的图像往往存在各种缺陷：摄像头传感器可能引入噪点（如夜间成像的雪花点），拍摄角度可能导致图像畸变（如手机斜拍高楼时的“透视变形”），光线不足则可能让图像模糊昏暗。预处理层的作用就是通过技术手段修复这些缺陷，输出高质量、规范化的图像，为后续特征提取“扫清障碍”。这一环节相当于对“原始素材”进行筛选和优化，直接影响后续处理的效率与精度。

预处理层的核心技术包括以下三类。

图像去噪：针对传感器或传输过程中引入的噪声（如 CMOS 传感器的热噪声、低光环境下的颗粒噪声），通过算法过滤干扰。常用方法包括高斯滤波（平滑图像去除高频噪声）、中值滤波（有效去除椒盐噪声，如老照片上的斑点）、非局部均值去噪（通过寻找相似像素块修复噪声区域，同时保留细节）。例如，在医疗影像场景中，X 光片的噪声可能掩盖病灶，去噪处理能让医生更清晰地观察骨骼结构。

图像增强：通过调整图像的亮度、对比度、色彩等参数，提升关键信息的辨识度。例如，在监控录像中，夜间图像往往偏暗，可通过“直方图均衡化”扩展灰度范围，让暗部细节（如行人轮廓）显现；在遥感图像中，可通过“色彩映射”增强植被（绿色）与裸地（褐色）的差异，便于后续分类。

几何校正：修正因拍摄角度、设备误差导致的图像畸变。例如，用手机俯拍桌面时，矩形的书本可能因透视关系变成梯形，“透视校正”算法可将其还原为标准矩形；工业相机拍摄流水线上的零件时，可能因镜头光学特性产生“鱼眼效应”，“畸变校正”算法可通过预设的镜头参数模型修正偏差。

预处理的最终目标是“保留有用信息、去除干扰信息”。例如，在人脸识别场景中，预处理会先裁剪掉图像中的背景区域（只保留人脸），再通过灰度化（减少色彩干扰）、光照归一化（消除强光或阴影的影响），为后续的特征提取提供“干净、标准”的输入。

3. 特征提取层：从图像到“关键信息”的“提炼器”

特征提取层是计算机视觉的“核心引擎”，其任务是从预处理后的图像中提取能表征物

体或场景本质的“关键特征”——这些特征是计算机理解图像的“语言”，也是区分不同目标的核心依据。如果将图像比作一篇文章，特征提取就相当于“提取关键词”，让系统能快速抓住核心信息。

特征提取的关键在于“找到具有区分性的信息”。例如，区分“猫”和“狗”时，耳朵形状（猫耳尖、狗耳圆）、尾巴形态（猫尾细长、狗尾粗短）就是有效特征；区分“苹果”和“橘子”时，颜色（红/绿 vs 橙黄）、纹理（光滑 vs 略带颗粒）就是关键特征。常用的特征提取技术可分为以下两类。

（1）传统手工特征：由人工设计算法提取特定特征，适用于简单场景。

① 边缘检测（如 Canny 算法）：通过识别图像中灰度突变的区域（如物体轮廓），提取“边缘特征”——这是区分物体形状的基础，如圆形的苹果和方形的盒子可通过边缘形状快速区分。

② 纹理分析（如 LBP 算法）：通过统计像素灰度的分布规律（如粗糙、光滑、条纹状），提取“纹理特征”——例如，树皮的粗糙纹理与玻璃的光滑纹理可通过纹理特征有效区分。

③ 颜色特征（如 RGB 直方图）：通过统计图像中红、绿、蓝三通道的像素分布，提取“颜色特征”——例如，在农产品分拣中，可通过颜色特征区分成熟（红色）与未成熟（绿色）的草莓。

（2）深度学习特征：通过神经网络自动学习特征，适用于复杂场景。在深度学习中，特征提取是“端到端”的过程：底层网络自动提取边缘、纹理等基础特征，中层网络将基础特征组合成更复杂的特征（如“眼睛＋鼻子＋嘴巴”组成人脸部件），高层网络进一步提炼出抽象特征（如“特定人的面部特征”）。这种方法的优势是无须人工设计特征，能自动适应复杂场景（如识别不同姿态、不同光照下的同一物体）。例如，在自动驾驶中，深度学习特征可同时提取车辆的轮廓、车牌、车灯等多维度信息，大幅提升识别的稳健性。

特征提取的质量直接决定后续任务的效果。例如，在手写数字识别中，如果能准确提取“数字的笔画走向”特征，即使数字写得潦草，系统也能正确识别；反之，如果提取的特征是“纸张的褶皱”（无关特征），则会导致识别错误。

4. 高层理解层：从特征到“语义认知”的“决策者”

高层理解层是计算机视觉的“最终输出端”，它基于特征提取层得到的关键特征，通过分析与推理实现对图像或视频的语义理解，最终输出人类可理解的结果（如“这是一只猫”“前方有行人横穿马路”）。这一环节相当于系统的“大脑”，将抽象的特征转换为具体的认知。

高层理解层的任务可分为基础任务与复杂任务：

基础任务：针对单一目标或简单场景的理解，如物体识别（判断“这是什么物体”）、目标检测（定位“物体在图像中的位置”并标记）、图像分类（将图像归为预设类别，如“猫”“狗”“汽车”）。例如，在电商平台的“拍照搜物”功能中，系统先提取商品的特征（如形状、图案），再通过特征匹配识别出“这是一款运动鞋”，并返回相关商品链接。

复杂任务：针对多目标、动态场景的综合理解，如场景分割（将图像按语义划分为不同区域，如“道路”“建筑”“天空”）、行为分析（判断视频中人物的动作，如“跑步”“举手”）、事件预测（基于当前场景推断未来可能发生的事件）。例如，在安防监控中，系统通过分析人群的移动特征（如聚集、奔跑），可判断“是否发生异常事件”；在体育赛事直播中，系统通过识别运

动员的动作特征(如“投篮”“传球”),可自动生成比赛精彩瞬间集锦。

高层理解的核心是“推理能力”,这一能力通常通过两种方式实现:传统方法依赖预设的规则(如“如果检测到红色信号灯且车辆在停止线内,则判断为‘等待通行’”);深度学习方法则通过训练数据学习“特征与语义的关联”(如大量“猫”的特征输入后,系统会自动记住“猫的特征组合对应‘猫’的标签”)。随着技术发展,高层理解正从“单一任务”向“多任务融合”发展——例如,自动驾驶系统需同时完成“检测行人”“识别交通灯”“判断车距”等任务,并综合这些信息做出“减速”或“通行”的决策。

5. 各层协同:从“数据”到“认知”的完整链路

计算机视觉的技术体系架构是一个“层层递进、相互依赖”的有机整体:数据输入层提供“原材料”,预处理层提升“原材料质量”,特征提取层提炼“核心信息”,高层理解层输出“最终认知”。任一环节的缺陷都会影响整体效果——例如,数据输入层采集的图像模糊度过高(如摄像头失焦),即使后续预处理和特征提取算法再优秀,也难以准确识别物体;反之,如果特征提取层未能抓住关键特征(如将“鸟的翅膀”误判为“飞机的机翼”),高层理解层必然会得出错误结论。

以“外卖自动分拣”场景为例,这一架构的协同过程清晰可见:

数据输入层:工业相机拍摄传送带上的外卖餐盒,生成彩色图像;

预处理层:去除图像中的传送带纹理噪声,校正因拍摄角度导致的餐盒变形;

特征提取层:提取餐盒上的订单编号(文字特征)、目的地标签(颜色特征);

高层理解层:识别订单编号对应的配送区域,判断“餐盒应被分拣至 A 区域传送带”,并向机械臂发送指令。

这一过程中,4 个层级紧密配合,最终实现了“从图像采集到智能决策”的全流程自动化。理解这一架构,有助于我们把握计算机视觉技术的核心逻辑——无论应用场景如何变化,其本质都是通过这 4 个环节的协同,将物理世界的光学信号转换为可用于决策的语义信息。

5.1.4 图像与视频的数字化表示

图像的数字化是将连续的图像信号转换为离散的数字信号的过程,主要涉及像素和颜色模型两个核心概念。像素是图像的基本单位,一张图像由大量像素组成,每个像素包含颜色和亮度信息。常见的颜色模型有 RGB(红、绿、蓝)模型,它通过三种基色的不同组合表示各种颜色(图 5-1);还有 HSV(色调、饱和度、明度)模型,更符合人类对颜色的感知习惯。

视频则是由一系列连续的图像(帧)组成,视频帧是视频的基本组成单元,如图 5-2 所示,每帧都是一个静止的画面,帧与帧之间存在时间上的关联性。视频的数字化除了包含每帧图像的数字化信息外,还需要记录帧率(单位时间内的帧数)、分辨率等参数。视频帧率是指每秒显示的帧数,通常用每秒帧数(Frames Per Second,FPS)或“赫兹”(Hz)来表示。视频帧率对画面的流畅度有重大影响。较高的帧率可以提供更流畅、更逼真的动画效果。图 5-3 是视频帧率的例子,帧率影响视频的流畅度,一般来说帧率越高,视频画面越流畅。一般而言,帧率达到 24 帧/秒时,人眼会感知到连续的动态画面。

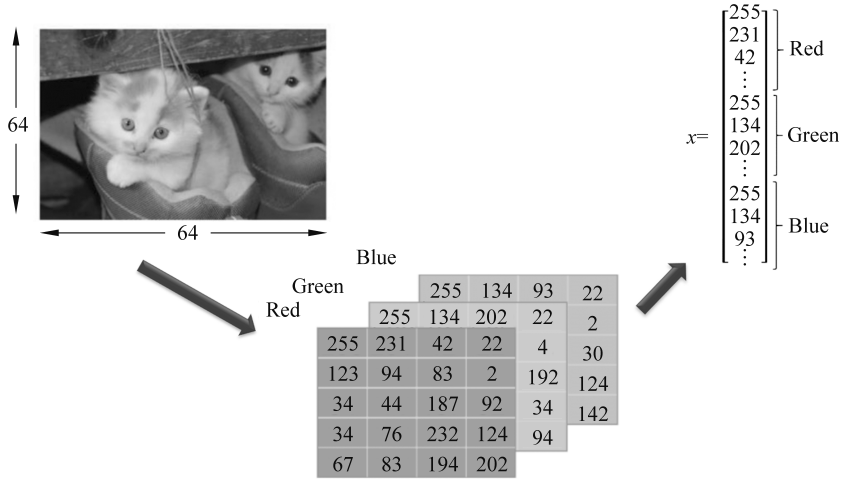

图 5-1　图像的数字化表示

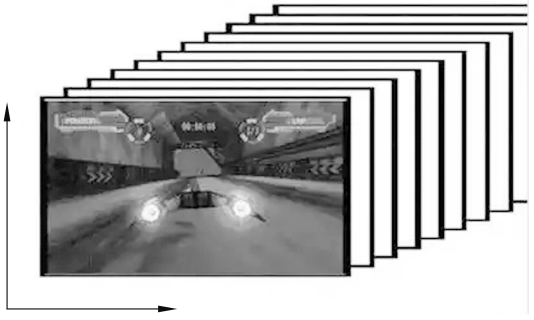

图 5-2　视频帧

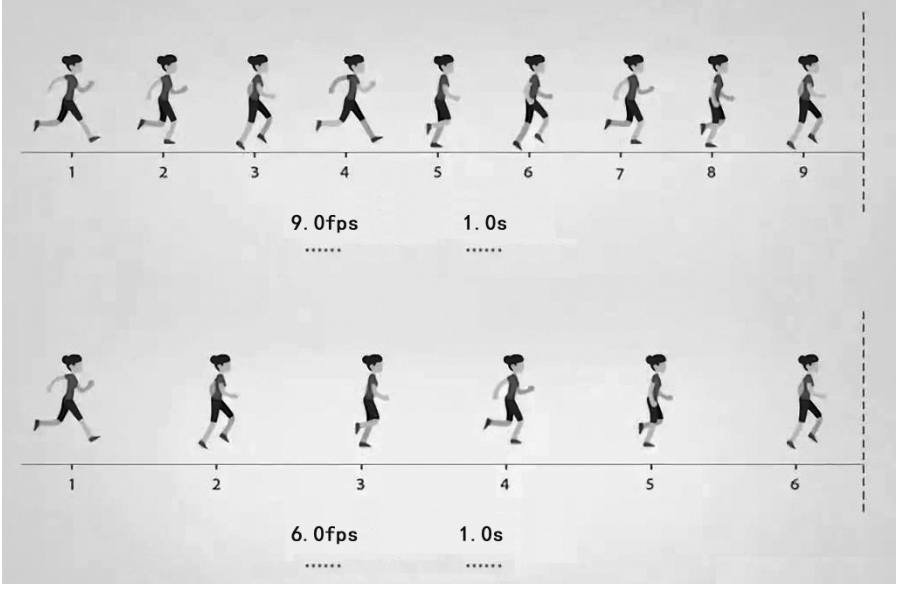

图 5-3　视频帧率

5.2 图像识别与分类

5.2.1 图像识别的基本流程

图像识别是计算机视觉中最基础也最核心的任务之一，其本质是让计算机通过对图像的分析，判断“这是什么”。这一过程并非简单的“看图识物”，而是一套标准化的技术流程——从原始图像采集到最终输出识别结果，需经过图像采集、预处理、特征提取、模型训练和分类识别五个紧密衔接的步骤。每个步骤如同链条上的一环，前一步的质量直接影响后续环节的效果，最终共同决定识别的准确率。

1. 图像采集：为识别提供“原始素材”

图像采集是整个流程的起点，指通过摄像头、扫描仪等设备获取目标对象的图像数据。这一步的核心要求是保证图像的“有效性”，即清晰记录目标的关键特征，避免因采集不当导致后续识别失败。

具体来说，需关注三个关键点：一是清晰度，例如在二维码识别中，模糊的图像会导致编码信息丢失，必须保证镜头对焦准确；二是完整性，在零件缺陷识别场景中，如果拍摄角度偏差导致零件边缘被截断，可能遗漏裂纹等关键特征；三是场景适配，例如识别透明包装的商品时，需避免强光反射导致的光斑遮挡——这些细节直接决定了原始数据是否“可用”。

简单来说，图像采集就像“为识别任务拍一张合格的‘证件照’”，只有素材合格，后续步骤才有意义。

2. 预处理：让图像“符合分析标准”

采集到的原始图像往往存在“个性化差异”：例如手机拍摄的图像尺寸可能是 1080×1920，相机拍摄的可能是 2048×2048；有的图像偏亮，有的偏暗；背景中可能混入无关物体（如识别书本时拍到的桌面杂物）。预处理的作用就是通过标准化操作，消除这些差异，让图像“整齐划一”地进入后续环节。

常见的预处理操作如下。

尺寸调整：将不同分辨率的图像统一缩放至固定尺寸（如 224×224），避免因大小差异干扰特征提取。

色彩归一化：通过调整亮度、对比度或转换为灰度图，消除光线变化对颜色特征的影响（如将阴天拍摄的水果图像与晴天拍摄的水果图像的色彩基调统一）。

干扰去除：通过背景分割算法剔除无关元素（如识别手写数字时去除纸张的褶皱），或用去噪算法过滤传感器引入的斑点噪声。

预处理相当于“给原始图像做标准化处理”，让所有待分析的图像处于“同一起跑线”，为后续步骤降低难度。

3. 特征提取：找到区分目标的“关键标记”

经过预处理的图像虽然规整，但仍包含大量冗余信息（如识别苹果时，果皮上的细小纹路可能无关紧要）。特征提取的任务就是从这些信息中“抽丝剥茧”，提炼出能区分不同类别的“关键特征”——这些特征是目标的“身份标记”，也是后续识别的核心依据。

例如，区分“苹果”和“橙子”时，颜色（红/绿 vs 橙黄）和形状（略扁圆形 vs 近圆形）是有

效特征；区分“猫”和“狗”时，耳朵形状（尖耳 vs 圆耳）和毛发纹理（短毛 vs 长毛）是关键差异。特征可以是简单的单一特征（如颜色），也可以是复杂的组合特征（如“圆形轮廓＋红色＋光滑表面”组合对应苹果）。

在技术实现上，既可以通过传统算法手动设计特征（如用边缘检测提取形状），也可以通过深度学习自动学习特征（如神经网络自主发现“苹果蒂的弯曲弧度”这类人眼不易察觉的特征）。无论哪种方式，核心目标都是找到“能让计算机快速区分‘是 A 还是 B’”的标志性信息。

4. 模型训练：让模型“学会识别规律”

特征提取得到了目标的“关键标记”，但计算机还不知道“这些标记对应什么对象”——模型训练就是让计算机通过大量样本“学习规律”的过程。这就像教孩子认识动物：需要先展示几十张猫的图片，告诉它“这些都叫猫，它们有尖耳朵、长尾巴”；再展示几十张狗的图片，告诉它“这些都叫狗，它们有圆耳朵、短尾巴”，孩子才能逐渐掌握区分规律。

模型训练需依赖“标注好的样本数据”（即已知类别的图像）：例如训练“花卉识别模型”时，需要准备数千张标注为“玫瑰”“百合”“郁金香”的图像，每张图像都附带对应的类别标签。训练过程中，模型会不断对比“自己的判断”与“正确标签”的差异，逐步调整内部参数（如调整“玫瑰的花瓣边缘锯齿更明显”这一规律的权重），直到能准确根据特征判断类别。

训练的核心是“数据量与质量”——样本越多、覆盖的场景越全（如不同角度、不同光照下的玫瑰），模型学到的规律就越通用，后续识别未知图像时才会更可靠。

5. 分类识别：输出最终的“识别结果”

经过训练的模型已经“掌握了识别规律”，分类识别就是让它“实际应用”的环节。当一张新的待识别图像输入后，模型会先通过预处理和特征提取得到其关键特征，再将这些特征与训练时学到的规律对比。例如，如果特征符合“尖耳朵、长尾巴”的规律，就判断为“猫”；如果符合“圆耳朵、短尾巴”，则判断为“狗”，最终输出明确的类别标签。

这一步的效果取决于前序环节的积累：如果图像采集清晰、预处理到位，特征提取就能抓住关键信息；如果模型训练充分，学到的规律就足够稳健——这些条件共同保障了最终识别结果的准确性。例如，在超市自助结账的商品识别中，系统通过上述流程，能快速将摄像头拍摄的商品图像归类为“可乐”“面包”等，并自动结算价格，这正是图像识别流程的典型应用。

5.2.2　传统图像特征提取方法

在计算机视觉发展的早期阶段，传统图像特征提取方法发挥了核心作用。这些方法主要基于图像的低层视觉信息（如边缘、纹理、颜色等）进行人工设计，具有明确的物理意义和可解释性。表 5-2 介绍了以下几种经典方法。

（1）边缘检测（Edge Detection）。

原理：通过识别图像中像素灰度值发生显著变化的区域（梯度极大值点）来定位物体的轮廓。

常用算子：Sobel 算子、Canny 算子等。

应用示例：在车牌识别系统中，边缘检测可快速定位车牌的矩形边框。

（2）纹理特征提取（Texture Feature Extraction）。

原理：纹理描述了图像局部区域中像素灰度或颜色的空间分布模式（如重复性、方向