# 前　言

情感识别，作为人工智能领域的一个重要分支，近年来在学术界和工业界都受到了广泛的关注。随着深度学习、计算机视觉、自然语言处理等技术的不断进步，情感识别的应用场景也日益丰富，从基础的人机交互、智能客服，到高级的心理健康监测、情感智能分析等，情感识别技术正逐步渗透到我们生活的方方面面。

本书所探讨的内容正是围绕情感识别的最新研究进展和关键技术展开的。通过对大量文献资料的梳理和总结，本书力求为读者呈现一个全面、深入且前沿的情感识别知识体系。

在情感识别的众多技术手段中，深度学习无疑占据了举足轻重的地位。深度学习模型以其强大的特征学习能力和泛化能力，在图像、语音、文本等多种模态的情感识别任务中均取得了显著的效果。本书详细介绍深度学习在情感识别中的应用，包括卷积神经网络、循环神经网络、图卷积网络等主流模型的优缺点及在情感识别任务中的具体实现方法。

在图像情感识别方面，深度学习模型能够自动从人脸表情、身体姿态等视觉特征中提取出与情感相关的信息，进而实现对情感状态的准确判断。本书不仅介绍传统的基于人脸表情识别的情感识别方法，还探讨如何结合上下文信息、情境描述等多模态信息来提升识别的准确性和鲁棒性。此外，对于视频情感识别这一更具挑战性的任务，本书也给出基于深度学习的解决方案，包括如何有效地处理视频序列中的时间依赖性和空间相关性等问题。

在语音情感识别方面，深度学习同样展现出了强大的优势。语音作为人类交流的重要工具，其中蕴含着丰富的情感信息。本书详细分析语音信号中的韵律特征、音质特征等，并介绍如何利用深度学习模型从这些特征中提取出与情感相关的信息。同时，本书还关注语音情感识别中的一些问题，如情感信息分布不均衡、噪声干扰等，并给出相应的解决方案。

除了图像和语音情感识别外，本书还探讨文本情感识别的相关技术。文本作为人类表达情感的一种重要方式，其情感识别任务同样具有重要意义。本书介绍基于自然语言处理的文本情感识别方法，包括情感词典、机器学习模型和深度学习模型等，特别是结合了情感常识知识库和外部知识信息的文本情感识别方法，能够更准确地理解文本中的情感信息，为情感智能分析提供了有力的支持。

此外，本书还关注情感识别的多模态融合问题。在实际应用中，情感信息往

往通过多种模态进行表达和传播。因此，如何将多种模态的情感信息进行有效的融合和利用，成为一个亟待解决的问题。本书介绍多模态情感识别的基本原理和方法，包括多模态特征提取、特征融合和决策融合等关键步骤，并给出具体的实现案例和分析结果。

本书是在作者研究团队多年研究成果的基础上撰写完成的，其中王科俊负责统筹全书，陈静为本书的撰写和修改作出主要贡献，卢桂萍绘制了本书部分图形，并与方宇杰参与了部分撰写和校稿工作，杨涛等对本书的相关成果也作出了重要贡献。本书通过对深度学习、计算机视觉、自然语言处理等多种技术手段的综合运用和深入探讨，力求为情感识别的研究和应用提供新的思路和方法。

希望本书能够为广大读者在情感识别的学习和研究中提供有益的帮助和启示。同时，也期待更多的研究者和开发者能够加入情感识别的研究队伍中来，共同推动情感智能技术的不断进步和发展。

王科俊

2026 年 2 月

# 目　录

# 第1章 绪 论

人工智能先驱 Minsky 在 *The Society of Mind* 书中说道，问题不在于智能机器是否能拥有情感，而在于机器实现智能时怎么能够没有情感[1]。尽管现有技术在对象分类[2]、图像分割[3]等智能感知任务中取得了重大突破，但真正智能系统的建设仍然处于起步阶段，其中一个重要的原因是情感计算发展的限制，如小爱音箱、天猫精灵等智能音箱，它们能识别用户的指令，但并不能有效地识别用户表达的情绪。如果用户使用不同的情绪表达相同的内容，智能音箱的反应可能会是相同的。因此，智能机器在研发过程中需要加入情感计算的技术手段，才能变得更智能和人性化。

众所周知，情感计算是一个多学科的交叉任务，它涉及计算机科学、心理学、认知科学等领域。虽然它不会像人脸识别、行人再识别等研究那样具有独立的应用空间，但情感计算能够融入很多领域，潜移默化地影响和改善人们的生活，如笑脸抓拍在手机相机中的应用。疲劳驾驶检测[4, 5]可以通过收集驾驶员的面部和语音等数据进行分析，以确定驾驶员是否出现了疲劳驾驶并及时反馈给后台人员。疲劳驾驶检测在公用运营管理中尤为重要，它可以在很大程度上降低疲劳驾驶带来的事故。在医疗领域，痛苦检测[6,7]可以帮助医护人员及时发现患者的突发状况。在心理咨询和治疗方面，情感分析助手可以帮助医生识别和跟踪咨询人或患者的情感状态，及时调整治疗方案[8,9]。尤其是对于孤独症和社交恐惧的患者，非人类的咨询治疗有助于他们更好地敞开心扉[10,11]。而且辅助治疗手段也能缓解心理咨询对于心理医生本人的情绪和心理影响。由此可见，情感计算在辅助心理治疗方面有很大的应用价值。在零售行业，情感计算也有较为丰富的应用，包括智能客服、广告优化和店员服务质量管理等。智能客服[12,13]能够对用户的问题给予及时的反馈，并在用户表达消极情绪时及时转到人工服务，既避免了用户排队等待时产生焦虑感等负面情绪，又降低了人工成本，同时提高了服务质量。广告优化通过辨识顾客的个人属性信息，切换不同的广告内容，同时检测顾客的情绪状态及时反馈投放效果。在人机交互方面[12,14]，机器可以通过检测并识别用户的情绪做出对应的反馈，使机器做到有“温度”。情感计算必将使未来的家政服务机器人在完成指令的同时更加人性化，在一定程度上起到陪伴的作用，进而提高独居老人的生活质量，减少独居老人的孤独感。随着情感计算的发展，尤其是在自然环境下的情感识别系统的开发，使其在上述领域的商业应用方面展现了巨大的潜力。

然而，让机器准确理解人类的情感并不是一件容易的事。人类通过长时间的经验知识积累能轻松地感知到他人的情绪状态和情感变化，但对机器而言，情感是非常抽象和复杂的。虽然在心理学上对情感的定义依旧存在争议，但是不同研究者仍会结合心理学的研究从不同的角度对情感进行描述。对于情感的感知需要结合具体的情境、人物和场景等信息，而且在对情感量化和识别的过程中需要对不同的信息来源和影响因素进行取舍。这些复杂的信息和影响因素导致的不确定性使机器对人类情感的认知和理解非常具有挑战性。从 20 世纪 90 年代开始就有不少研究机构陆续开展了情感计算相关的工作，如国外的麻省理工学院、奥卢大学、卡耐基·梅隆大学等，国内的中国科学院自动化研究所、清华大学、东南大学等。这些机构的研究人员从不同的角度分析了情感，不断拓展情感分析的研究内容和方向。近年来又涌现了一大批以情感计算为核心的公司，如 Affectiva、阅面科技、Kairos 等。这标志着情感计算的研究成果正逐步走向实际应用。情感计算在现实中的应用又为它在科研领域的发展提供了新的研究思路并带来了更大的挑战。例如，如何在保证识别性能的同时降低对硬件的需求，以实现实时的情感追踪，同时还需要适用各种不同的场景，解决真实环境中的遮挡、光照等问题。另外，为了更准确地理解情感，还需要对情感发生的前因后果等信息进行提取和分析。综上所述，情感计算作为当前人工智能领域的热点研究问题之一，具有重要的理论研究意义和实际应用价值，与此同时也存在很多亟待解决的难题。

本书将自然环境下的图像和视频作为研究对象，目的是构建能在真实世界中自动提取有效情感特征的模型用于情感状态识别，从而满足在复杂且多变环境下人类情感捕捉的需求。此外，情感计算正处于蓬勃发展的阶段，利用易于采集的外在情感信息(面部表情、声音、身体姿态等)综合分析判断情感状态是其中一项重要的研究工作。本书在这些模态的基础上引入了更宽时间范畴的情境信息，在一定程度上赋予了情感识别系统先验信息，有利于更广泛场景下的多模态情感计算。综上所述，本书的研究内容具有重要的理论意义和应用前景。

## 1.1　国内外研究现状概述

情感计算的概念最早由麻省理工学院的 Rosalind W. Picard 在 1995 年提出。在 1997 年，Picard[15]出版了情感计算相关专著。情感计算的研究目的是使机器能够模拟和计算人类的情感。目前情感计算的技术有很多种，包括以人为主体的面部表情识别、语音情感识别、肢体语言情感识别、身体数据监控识别，以及自然语言处理领域的文本情感分析、情感极性识别等。虽然情感计算的概念是在 1995 年提出的，但一些子任务的研究要早于这个时间节点。早在 1978 年，Sown 等[16]

就进行了面部表情识别的最初尝试。1991 年，Mase[17]提出了使用光流法进行面部表情识别的方法，至此自动人脸表情识别研究进入了一个新的时期。在深度学习兴起之前，人脸表情识别的方法主要是 Gabor 小波[18]、神经网络[19]、光流法、隐马尔可夫模型[20]、几何追踪法等。身体姿态情感识别最初的方法有外观　　法[21]和三维建模法[22]。这些传统方法使用的数据是用电磁感应器和光学反射标志等辅助设备，在简单背景、简单身体关节移动方向等约束条件下收集的。语音情感识别则是使用基音变量、语速等韵律特征，结合 $K$ 近邻、支持向量机（support vector machine，SVM）等机器学习的方法进行情感识别[23,24]。2004 年，Coan 等[25]在研究中使用了脑电波计算心情。随着单一模态的情感识别技术的发展，有些工作也开始尝试使用多种模态的信号进行多模态情感识别[26,27]。以上这些传统的情感识别研究都基于可控的实验室环境下采集的数据集，要求被采集人员做相应的表情，然而这并不是真实情感的表达，而且实验室环境的背景简单、光照无变化、参与实验人员少。这些都导致基于这类数据集设计的情感识别系统实际应用效果不理想。相较于这类研究，自然环境下的情感识别方法的研究更契合人机交互等领域的实际应用需要。得益于越来越多自然环境数据集的建立，自然环境下情感识别方法的性能逐步提高。而深度学习的兴起也使情感识别的研究进入了新的时代。因此，接下来主要介绍基于深度学习方法的情感识别研究工作。

本书的研究是基于图像和视频且以人为主体的，所以本节主要介绍以人为主体的静态图像和视频序列情感识别研究现状。其中，基于静态图像的情感识别相关研究主要围绕以下方向展开：解决自然环境中困难样本引发的问题，融合深度学习与传统方法，结合联合辅助任务开展面部表情识别，应对光照变化、遮挡及姿态变化等问题，以及基于身份解耦合、场景上下文信息的研究。基于视频的情感识别主要围绕单模态的序列建模和多模态的情感识别研究进行介绍。在本节的最后还介绍自然环境下的情感数据集。

### 1.1.1　基于静态图像的情感识别研究现状

自然环境下的数据集存在许多困难的样本，其中包括各种程度的遮挡或分辨率低的图像。而且情感数据标注本身就带有主观性，因此数据集也会包含很多噪声标签。为了缓解标签噪声[28]和困难样本带来的影响，Gan 等[29]提出了使用软标签来替代传统的硬标签的方法。在训练阶段硬标签被用作监督信息训练模型得到各类别的预测概率，然后通过融合各子集中的预测概率得到软标签。接着，将软标签作为监督信息训练多个卷积神经网络（convolutional neural network, CNN）分类器，并在预测阶段集成多个 CNN 基分类器的结果。Zeng 等[30]提出了一个潜在真理的不一致伪标注框架，用它挖掘真实标签来解决标签不一致性问题。该框架

分为三个步骤：第一步使用两个人类标注的数据集训练深度模型；在第二步中交换数据集，用训练好的模型预测另一个数据集得到预测的伪标签，此外还引入了未被标记的数据集，同样应用两个训练好的模型给出预测的伪标签；第三步中先用 Dawid 等[31]提出的最大期望算法（expectation-maximization，EM）估计潜在的真实标签，然后使用估计的标签训练整个 LTNet。Wang 等[32]提出了一个自修复网络来抑制标签的不确定性。首先使用损失重新加权的自注意力模块确定每个批次中样本的重要程度。然后按重要性进行降序排序并划分成两组，同时引入了秩正则化损失来约束重要性分组的差异。之后修改了低重要性组中的一些不确定性样本的标签，最后利用修改后的标签重新训练了网络模型。以上方法都是通过获取数据的伪标签，利用设计的规则判断是否更新标签，并利用更新的标签重新训练模型。与上述方法不同，文献[33]和[34]提出了一种协同协作训练框架，利用监督损失和一致性损失的凸组合来联合训练。监督损失促使网络逐步学习面部情感特征，而一致性损失通过强制多个网络保持一致来抑制噪声标签的影响。随着训练的进行由监督损失逐步转移到一致性损失，既保证了对干净数据的充分利用，又使任何网络本身不对嘈杂标签过度拟合。

除了单独使用深度网络提取特征外，一些研究也采用了传统的手工提取特征与深度特征相结合的方法。Otberdout 等[35]使用协方差描述符对局部和全局的人脸特征进行了编码，并利用 SVM 与定义的正定高斯核对特征进行分类。该方法首先利用深度网络提取出特征，然后使用手工描述符进行编码。而有些研究是先利用手工描述符提取图像的特征，然后输入深度网络模型中进一步完成情感特征提取和识别。Mohan 等[36]使用局部引力描述子[37]提取了图像幅度（$M$）和方向（$D$）上的向量，其中 $M$ 表示边缘强度的信息，$D$ 表示图像的几何结构。接着将 $M$ 和 $D$ 分别输入深度卷积神经网络（deep convolutional neural network，DCNN）中进行特征提取，并在最后进行特征融合。文献[38]使用了两个 Gabor 滤波器提取图像特征，然后将特征图输入 CNN 中。文献[39]使用了一个多分支模型，其中三个分支使用了预训练的视觉几何组（visual geometry group，VGG）模型，另一个分支利用尺度不变特征转换（scale-invariant feature transform，SIFT）描述符[40]提取图像的密集特征，然后利用 $K$ 均值聚类[41]的方法将描述符转换为视觉特征来构建特征表示，最后将这四个分支提取到的特征进行拼接融合。除了以上两种结合方式，特征描述符还可以融合到深度学习的基础模块中。Jiang 等[42]分析了 Gabor 滤波器[43]提取视觉特征的优势，并利用它设计了一个最优的 Gabor 卷积网络。尽管深度学习的方法取得了不错的效果，但缺乏可解释性，因此还有不少学者仍然坚持传统算法理论的研究。针对表情识别任务，Iqbal 等[44]提出了一种邻域感知边缘方向模式的局部描述符来解决由于噪声而存在弱边缘和扭曲边缘的问题。但他们的实验仅在可控条件

下收集的数据集上进行了测试，在真实环境数据上的效果仍有待考证。

表情识别和面部运动单元检测是两个相关性很高的任务，而基于面部的其他任务，如人脸识别、性别和年龄估计也都可以作为情感识别的辅助任务。接下来介绍与表情识别相关的多任务和多阶段工作。Wang 等[45]提出了一个多任务自适应权重共享网络。他们使用了两个 VGG19[46]作为主干网络进行特征提取，在每个池化层后放置一个自适应共享单元来桥接用于两个任务的单独网络，实现网络间信息的共享。文献[47]提出了一种两级注意力和两阶段多任务学习的情感识别框架。首先图像被输入两级注意力层，第一级注意力层使用残差注意力模块[48]提取不同感受野的位置特征，并将它们输入由卷积、平均池化和具有自注意力的双向循环神经网络（bidirectional recurrent neural network，BRNN）组成的第二级注意力层，以提取不同感受野间的关系信息。然后将融合了不同层级信息的特征输入多任务学习网络中。在第一阶段的多任务学习中，学习了离散分类任务和连续任务。在第二阶段的多任务学习中，将离散特征和连续特征拼接，用于同时预测价效值和唤醒值。Jin 等[49]提出了一个简单的多模态多任务模型，将图像和音频数据分别输入对应的特征提取网络中，得到的特征进行再融合并完成分类。在多任务中共用一个骨干模型提取特征，在训练时交替计算其中一个任务的损失并更新网络模型。该模型虽然简单，但效果不错，在 ICCV 2021 第二届“野外情感行为分析”研讨会与挑战赛（The 2nd Workshop and Competition on Affective Behavior Analysis in-the-wild，ABAW2）中获得了第二名的成绩。还有很多研究也是使用一个骨干网络，设计多个分类器完成不同的任务[50-53]。文献[54]提出了一种深度监督注意网络（deeply-supervised attention network，DSAN）来进行面部情感识别，并采用了两阶段训练法，在第一个阶段以种族、性别和年龄等与人口统计相关的信息为标签训练了 DSAN，在第二个阶段中固定了 DSAN 中的深度监督模块的参数，并在深度监督模块之后添加了新的注意力模块和全连接层，这样可以充分地利用学习到的人口统计学特征，并将它作为表情识别注意力模块的指导信息。

为了解决真实环境中的样本存在的光照变化、遮挡和姿态变化等问题，面部分区域和注意力的方法被广泛采用。Li 等[55]提出了具有注意力机制的 CNN（ACNN）用于构建遮挡感知面部情绪的情感识别系统。ACNN 中使用门单元将注意力从遮挡的面部部分转移到其他可见的面部区域。其中通过补丁门控单元为人脸特征图的多个子特征图赋予不同的权重，而全局门控单元将整个人脸的特征图编码为加权向量。加权后的局部和全局特征被拼接并分类。ACNN 中面部区域的分解是基于特征点的，因此它的性能取决于模型中人脸标志点定位模块的准确性。类似地，区域注意网络[56]是通过裁剪多个面部区域，并利用基于自注意力的模型来学习每个区域的重要权重。然而，基于自注意力的方法缺乏额外的监督来确保功能。因

此，在大遮挡和姿态变化下，网络可能无法准确定位这些未遮挡的面部区域。文献[57]提出了基于面部关键点引导的注意力分支来标记遮挡区域的特征。文献[58]使用元学习的方法在样本稀缺的情况下进行情感识别。首先提取支撑集的特征向量，当查询样本到达时使用相同的网络进行编码，然后计算查询样本和每类原型间的欧几里得距离，最后利用最近邻分类器进行分类。分区域的方法大多基于面部关键点，因此过度依赖人脸标志点定位的准确性，而且网络模型通常较大，参数量多且预测速度慢。

身份信息等干扰因素的存在，会影响对特定情感特征的提取。一些方法已经研究了如何实现情感识别的干扰解纠缠。Meng 等[59]提出了一种身份感知 CNN 方法，以减轻面部身份造成的差异。该方法利用身份敏感的对比损失来学习身份相关信息。Wang 等[60]提出了一种对抗性特征学习方法来解决姿态和身份带来的干扰。Yang 等[61]提出了一种身份自适应的方法来学习身份子空间，该方法可以在保留每个个体的身份相关信息的同时生成不同的表情，实现了身份特征与表情特征的解耦。Ruan 等[62]提出了一种自适应深度扰动解纠缠学习方法，首先训练了扰动特征提取模型来预测身份信息和姿态信息等干扰因素。然后在自适应解纠缠模型训练中，扰动子网络利用从扰动特征提取模型中学习到的特征，学习自适应扰动特有的特征。

除了利用面部图像进行表情识别外，情感图像内容分析也是一个研究领域。本书使用以人为主体的图像，因此这里不再介绍情感图像分析的内容，相关研究可查看文献[63]。

以往的情感识别工作大多基于面部图像或身体姿态图像，近年来的研究中有些学者开始探索使用场景语境信息。Kosti 等[64,65]创建了一个包含场景上下文信息的静态图像数据集，并提出了一个双流网络，其中一个分支将整幅图像作为输入学习场景上下文信息。Zhang 等[66]利用目标检测网络提取了若干候选框，并提取了每个框中的特征，然后将这些特征作为图卷积网络[67]的节点，通过节点间信息传递学习图像中各部分的关系。EmotiCon[68]是一个多模态融合的网络，其中两个分支被用来建模场景上下文信息：一个分支将主体遮挡后的图像输入特征提取网络中学习除主体外的场景信息；另一个分支则将图像处理为深度图，并使用了一个常规卷积网络学习全局特征。Thuseethan 等[69]将主体、非主体和整幅图像分别输入三个 CNN 中，最后通过拼接融合进行情感识别。

### 1.1.2　基于视频序列的情感识别研究现状

基于视频的情感识别方法通常是先利用CNN提取视频帧的特征，然后输入处理时序的网络（如长短期记忆（long short-term memory，LSTM）网络[70]或门控循环

单元(gated recurrent unit，GRU))中进行特征融合并输出分类结果[71]。Zhen 等[72]将 CNN 提取的图像特征进行张量化，然后将张量化的时间序列人脸特征输入 LSTM 网络中，以达到压缩网络规模的目的。文献[73]使用了受到广泛关注的 Transformer 模型来创建时空提取网络。视频序列首先通过 4 个卷积模块得到高层语义特征，然后与空间位置编码相加再送到多个多头的空间Transformer模块中，卷积空间 Transformer 模块提取的特征输入由若干个多头子注意力和前馈网络组成的时间 Transformer 模块中学习时间序列中的面部特征关系。除此之外，还有同时进行时空特征提取的方法。文献[74]提出了双流卷积网络，一个分支是静态流，另一个分支是动态流，并且两个分支的每个卷积层共享了特征，经过特征提取后的特征输入自注意力模块中，来寻找与情感相关的区域。以上方法都是直接将图像序列输入网络中，而 Lee 等[75]提出将 RGB 图像、深度图和热敏记录图像作为多模态模型的输入。首先将三种图像分别输入空间编码器中提取特征，然后在时间解码器中融合了多模态的信息并为每个位置生成了注意力权重，加权后的多模态特征被输入情感识别网络中。不同于之前显式处理时序信息的注意力，他们将时空编码器作为时空注意力模块，不仅提取了视频帧间的时间重要性，还给出了每帧空间像素级的注意力权重。文献[76]利用人脸标志点的相对位置和人脸局部区域的纹理特征来描述人脸动作编码系统的动作单元，并将它作为一个分支的输入来提取浅层特征，另一个分支使用卷积注意力从图像序列中提取深层特征。使用人脸关键点的方法更适用于无遮挡的情况，无约束条件下的数据可能存在无法提取关键点的情况，因此该方法有局限性。Kim 等[77]提出在空间特征提取阶段融入情感状态的方法，该方法的效果很大程度上取决于估计图像的情感程度算法，有一定的局限性；该方法仅在可控环境下收集的数据集上进行了实验，在自然状态下的情感数据集上的效果可能并不佳。文献[78]引入了 GCN 来学习特征图中更重要的位置，然后应用 LSTM 网络来学习 GCN 学习到的特征间的长期依赖关系。表情强度也被用来对 GCN 节点的特征进行加权。由此可见，表情强度可以作为辅助任务，帮助时序中的情感识别。Xia 等[79]先计算了各帧的重要性权重，然后提取每帧的特征。在空间特征提取时，他们使用了全局和局部特征提取结合的方法，将面部区域分成 9 部分，分别提取了特征并计算了每个区域的注意力权重。最后将全局和局部特征融合再进行分类。然而，他们提出的计算时间注意力的方法是基于 RGB 图像的，容易受到光照、遮挡的影响。上述方法都是在时空维度上使用了注意力机制，Chen 等[80]则将时空和通道注意力模块结合在一起，学习时空特征不同维度的重要性。

在 EmotiW 挑战赛中，很多队伍都使用了多个深度学习框架相结合的方法[81,82]，也确实取得了不错的效果。除了使用多个神经网络提取视觉特征外，还有很多研究融入了语音信号进行多模态的情感识别。一些工作提取了语音的声谱图，然后进行

特征提取，并与面部特征一起进行融合[83-85]。对语音信息的提取也可以采用一维卷积，将语音通过语音情感特征提取工具箱（如 OpenSMILE[86]）进行统计学上的特征提取，然后利用机器学习或深度学习的方法进一步提取特征[82,87-89]。除以上在各模态中提升特征表示能力的工作外，很多学者也研究了多模态融合的策略[84]。Zhou 等[84]对模态内和模态间的融合方法进行了探索。在模态内使用多种注意力突出关键的情感特征，然后使用因子双线性池化进行跨模态的特征融合。另外，还有一些工作开始研究除主体外的上下文信息的影响。Lee 等[90]提出了双流网络，一个分支用于提取面部的序列信息，另一个分支利用遮挡的面部图像序列来学习场景信息。Filntisis 等[91]使用了多种模态的输入，除身体和遮挡身体的图像序列外还使用了光流图，并探索了标签间的相关性并为此引入了嵌入损失。文献[92]在文献[91]的基础上进行了扩展，加入了其他几种输入：身体关键点、整幅图像、身体图像的光流图、除主体外的光流图和面部光流图。总之，上述方法对于场景信息的利用主要是将整幅图像作为输入或者将遮挡主体后的整幅图像作为输入。

### 1.1.3　语音情感识别国内外研究现状

语音情感识别系统通常包括特征提取和特征分类两个部分，如图 1.1 所示。第一部分通常包括语音的预处理和特征提取两个步骤，第二部分分为训练和测试两个阶段。在训练阶段，语音情感识别系统通常是构建分类模型并训练模型。在测试阶段，语音情感识别系统把待预测样本输入分类模型中，计算得出识别结果。

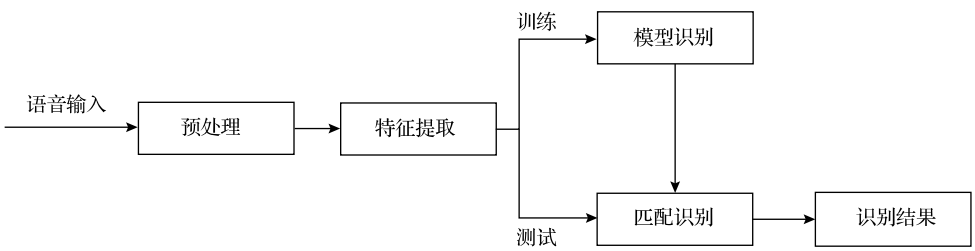


图 1.1　语音情感识别系统框图

早期，在特征提取这一部分，常用的特征大致可分三种：韵律特征[93]、基于谱的相关特征[94]和声音质量特征[95]。这三类声学特征通常以帧为单位提取，被视为局部特征，同时被称为低级描述符（low level descriptor，LLD）。在某些情况下，研究者会针对这三类特征计算它们的最大值、最小值、极值范围、均值、标准差等指标，并利用这些统计指标作为情感语音样本的全局特征参与分类器的训练。

常用的韵律特征包括共振峰、基频、能量等。2005 年，Luengo 等[96]首先对每句情感语音提取能量和基频曲线以及它们的对数形式，进而提取它们的一阶差分和二阶差分，一共得到了 12 条特征曲线，最后统计出每条曲线的最大值、最小

值、均值、方差、变化范围、偏斜度(skewness)、峰度(kurtosis)。每一个样本都对应 84 维韵律统计特征。对该特征集经过特征选择和实验验证后发现仅需要 6 个特征也能得到较好的结果，说明韵律特征对情感识别非常有用。这 6 个特征分别是基频均值、能量均值、基频方差、基频对数的斜交、基频对数的动态范围和能量对数的动态范围。Origlia 等[97]所采用的特征由基频及能量等相关特征的统计指标组成，是一个总计 31 个值的韵律特征集。

常用的基于谱的相关特征有线性谱特征和倒谱特征两种。Nwe 等[98]研究发现，各个频谱区间的能量强弱分布与该语音中的情感信息有着一定的关系。例如，“高兴”情感与高频段中的高能量具有一定的相关性，而“悲伤”情感在同样的频段却表现为相对较为明显的低能量。目前，基于谱的相关特征在与语音相关的任务上均有着成功的应用，其中就有基于谱相关特征的语音情感识别研究。因此，基于谱的相关特征有着较强的表征能力，不可被忽视。

常用的声音质量特征有共振峰频率及其带宽、频率微扰和振幅微扰、声门参数等。影响声音质量的情况有说话人哽咽、气喘吁吁等，这些通常伴随着情感的变化，因此声音质量特征与语音情感紧密相关。Lugger 等[99-101]将第一和第四共振峰频率及其带宽作为声音质量特征并与基频等韵律特征一起作为总特征集。Li 等[102]将频率微扰和振幅微扰作为声音质量参数，进行了说话人不相关的情感识别。

此外还有基于以上三种特征的融合特征，选择不同类型的特征组成有利于分类的特征集。Li 等[102]在柏林情感语料库(Berlin database of emotional speech，EMODB)和合成语料库的交叉数据库上使用过零率、能量、基频、声音质量、谐波噪声比、0～15 阶梅尔频率倒谱系数(Mel-scale frequency cepstral coefficient，MFCC)等特征的相关统计指标，共计 5967 维。

在分类器上，早期语音情感识别方法所使用的分类器包括高斯混合模型(Gaussian mixture model，GMM)、SVM、多层感知机(multi-layer perceptron，MLP)等。GMM 是一种用于密度估计的概率模型。SVM 中核函数的应用提高了数据的可分性，将线性不可分数据映射为新空间中的线性可分数据。MLP 具有通用非线性函数逼近器的性质，在模式识别问题中得到了广泛的应用。

2006 年，Hinton 等[103]指出含有多个隐含层节点的深度神经网络具有较强的特征学习能力，在训练上的问题可以通过逐层初始化来克服。2012 年，AlexNet[104]在 ILSVRC(ImageNet Large Scale Visual Recognition Challenge)图像分类任务中荣获桂冠，吸引了众多研究者的目光，引发了深度学习在各个方面的热潮。随着深度学习在语音情感识别中得到应用，语音情感识别的准确率得到了较大的提升，语音情感识别的网络结构也是层出不穷。现如今语音情感识别中的深度学习方法按照网络结构一般可分为 CNN、循环神经网络(recurrent neural network, RNN)、

CNN 和 RNN 的结合。按照所使用的 CNN 的维度可以分为一维模型和二维模型。一维模型使用一维 CNN（1D-CNN），通常针对 MFCC、韵律特征集等序列信号。二维模型使用二维 CNN（2D-CNN），通常针对频谱图等图像信号。此外，深度学习方法还可以根据特征提取和特征分类两个角度分为两大部分。

随着迁移学习在计算机视觉、文本分类、行为识别等领域的成功运用[105]，Lakomkin 等[106]采用迁移学习的方法构建识别模型，该模型并不是从头开始训练。该方法先获得一个已经训练好的语音识别网络，再使用其浅层结构对 MFCC 进一步提取特征后输入 GRU 网络中，结构如图 1.2 所示。目前语音识别已经有很多商业化的产品，这表明语音识别方法能够获得相当高的准确率和快速推理能力。因此，对于语音情感识别来说，其特征有一定的迁移价值。基准模型采用多层 GRU 网络从头开始训练。IEMOCAP 数据集的实验结果表明，改进后模型的准确率和 $F_1$ 分数相对于基准模型提高了 10%左右，准确率最高可达 61%。

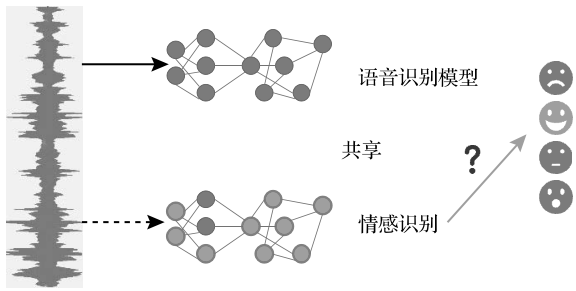


图 1.2　基于迁移学习的识别模型

自编码器（auto encoder）[107]是一种被广泛使用的无监督模型。语音情感识别中有标签的情感语音样本数量较少，这在一定程度上限制了深度学习模型的能力。因此，考虑到未标记语音的数据量，Deng 等[108]提出了半监督自编码器，通过标记数据和未标记数据的组合来提高语音情感识别准确率。该方法用未标记数据来训练自编码器，用自编码器提取的特征来表征样本，进而分类。进一步，该方法将模型的浅层特征与深层特征对应位置相加，这样的跳跃式连接有效地避免了梯度消失的问题。其输入特征是文献[109]中的 384 维特征集，由语音的 MFCC、过零率等特征的统计学参数构成。实验结果表明，所提出的方法能够在少量标记样本的情况下从未标记数据中学习知识，从而明显提高了识别性能，准确率达到 43.6%。

上述深度学习方法具有一个显而易见的特点，识别方法分为两步，侧重于如何通过深度学习模型表征样本，再进行分类。除此以外，语音情感识别中的深度学习方法通常使用 CNN 或 RNN 直接分类，更准确地说，针对声学信号这些低级描述符，使用 CNN 或 RNN 来进一步提取更高层次的话语级别特征并进行分类。