模式识别技术丛书

音视频情感识别的关键技术研究

张石清　赵小明　著

科学出版社

北　京

内 容 简 介

当前有关情感识别的研究已成为模式识别、人工智能、人机交互等领域中的一个热点研究课题，正越来越受到国内外科研机构和研究人员的重视。它在智能人机交互、机器人等领域具有重要的应用价值。本书共 10 章，首先介绍了当前面向语音和人脸的音视频情感识别技术研究概况，然后详述了音视频情感特征的提取及降维方法、音视频情感的分类方法、音视频情感信息的融合方法，最后重点介绍了基于近年来新发展起来的深度学习理论的音视频情感识别方法。本书大部分内容来源于作者在情感计算领域获得的多个省部级和国家级自然科学基金项目的资助下所取得的研究成果方面的归纳与总结。

本书可作为计算机类、电子信息类等专业的高年级本科生及研究生开展模式识别、人工智能、智能人机交互、情感计算等领域研究的参考教材，也可作为从事情感计算领域的科技工作者的参考书。

图书在版编目(CIP)数据

音视频情感识别的关键技术研究/张石清，赵小明著. —北京：科学出版社，2019.1

ISBN 978-7-03-058405-2

Ⅰ. ①音… Ⅱ. ①张… ②赵… Ⅲ. ①图像识别－研究 ②语音识别－研究 Ⅳ. ①TP391.413 ②H012

中国版本图书馆 CIP 数据核字（2018）第 171398 号

责任编辑：赵丽欣 常晓敏 / 责任校对：王万红

责任印制：吕春珉 / 封面设计：耕者设计工作室

科学出版社 出版

北京东黄城根北街 16 号

邮政编码：100717

http://www.sciencep.com

印刷

科学出版社发行 各地新华书店经销

*

2019 年 1 月第 一 版 开本：787×1092 1/16

2019 年 1 月第一次印刷 印张：8

字数：176 000

定价：79.00 元

（如有印装质量问题，我社负责调换（ ））

销售部电话 010-62136230 编辑部电话 010-62134021

前　言

情感在人与人之间的交流过程中起着非常重要的作用。利用情感表达所表现出来的外在信息，如人脸的面部表情或情感语音信号，人们就可以“察言观色”，了解别人的真实想法。这种情感能力是人类智能的一种重要标志。那么在人机交互的过程中，计算机能否像人一样，察觉用户的情感状态，如喜、怒、哀、乐，并且做出合适的反应呢？这就是情感计算研究领域要解决的一个重要问题。

情感识别是情感计算领域中极其重要的研究内容之一，它通过对采集到的用户情感信息进行分析和建模，然后识别出用户的情感状态。目前，有关情感识别的研究已成为模式识别、人工智能、人机交互等领域的热点研究课题，越来越受到国内外科研人员的重视。2017 年 7 月，在《国务院关于印发新一代人工智能发展规划的通知》中，感知智能是十分核心的研究领域，而情感识别则是感知智能技术极其重要的组成部分。本书作者近几年来在国家自然科学基金项目（61203257、61272261）、浙江省自然科学基金项目（Y1111058、LY16F020011）以及中国博士后基金项目（2016M591015）的支持下，开展面向语音和人脸表情的音视频情感识别方面的研究，包括音视频情感特征的提取及降维方法、音视频情感的分类方法、音视频情感信息的融合方法等。

本书是在这些项目的研究成果基础上，系统地加以归类总结撰写而成的。全书内容分为 10 章。第 1 章为绪论，简要介绍了情感识别的研究背景及应用。第 2 章介绍了基于语音和人脸的音视频情感识别技术回顾，包括语音情感识别和人脸表情识别中的特征提取与降维、分类方法以及深度学习技术在音视频情感识别中的应用。第 3 章为基于非线性流形学习的语音情感识别。第 4 章为融合核方法与非线性流形学习的语音情感识别。第 5 章为基于稀疏表示理论的鲁棒性人脸表情识别。第 6 章为基于特征层和决策层的音视频情感识别。第 7 章为基于 CNN 和 DTPM 的语音情感识别方法。第 8 章为融合 DBN 与 MLP 的人脸表情识别方法。第 9 章为基于多模 CNN 的音视频情感识别方法。第 10 章为基于混合深度学习的音视频情感识别方法。值得指出的是，第 7～10 章的内容包含作者近年来所提出的一些基于新发展起来的深度学习理论的音视频情感识别方面的创新研究成果。

由于作者水平所限，书中疏漏在所难免，敬请读者批评指正。

目　录

第 1 章　绪　　论

人非草本，孰能无情？每个人在人与人之间的交流过程中都会产生情感，不同的情感会对交流产生不同的影响。可见，情感在人类之间的交流过程中扮演着非常重要的角色，人们可以通过它的外在表现（如言语声调、面部表情、身体姿态等信息）交流彼此的想法、意图和愿望等。根据古文献的记载，人类很早就已经察觉到情感在交流过程中的重要性，甚至在古埃及时代的修辞著作中就已经对情感一词有所涉及。对于人类情感的研究一直是心理学、生理学和语言学等领域的重要方向，近年来开始受到工程研究领域的关注。其中一个出发点是辅助设计更加自然化、人性化的人机交互方式。但是，究竟什么是情感以及情感是如何表现、如何度量的？这些问题一直困扰着研究人员。本章将从这些问题出发，层层深入地叙述有关情感的定义及其表示理论、情感计算、语音情感识别、人脸表情识别、音视频情感识别的概念以及情感识别的应用。

1.1　情感的定义

什么是情感？这是一个很复杂的问题，从过去到现在一直是科学领域不断争论的一个问题。尽管心理学家已经对情感机理的研究做了大量的工作，但到目前为止，研究者对于情感的定义还没有达成一致的观点。在不同领域的研究中，研究者对情感的解释也不尽相同。

通常所说的情感，主要是作为一个心理学的概念而出现的，对它的定义，也多源自心理学。

我国《辞海》中对情感、情绪是从相近的意义上来进行定义的，即“情感”也称“感情”，是指人的喜、怒、哀、乐等心理表现。

杨泽民[1]认为，比较流行的情感的定义主要有 3 种：第一，所谓感情或情绪就是人对其所认识与处理一切的体验；第二，情感是人对客观事物的一种态度；第三，情感是人对于某一事物的态度的体验。然后，他提出一种情感的定义：情感是由非中性事物引起的并反作用于这个事物的非中性的意图和行为。

杨巍峰[2]认为，把情感定义为一种行为和意图，不能揭示情感的本质属性，因为没有反映出情感发生过程中机体生理生化变化的作用。根据情感发生过程的特点，即情绪和情感是在内外刺激所引起的皮层评估活动与受它影响而发生的机体变化的反馈信息，他将情感定义为：情感是人脑以主观体验形式，反映客观事物与主体需要关系的心理现象。

庞学铨[3]从哲学的角度分析后认为，德国哲学家 Hermann Schmitz 提出的新现象学情感理论更具有创新意义，即把情感理解成客观上把握到的一种具有空间性的力量、气氛。他认为，情感是不确定的宽度无限的气氛，情绪上震颤的人身体上通常可以感受到被嵌置于这气氛中。

美国心理学家 Willian James[4]和丹麦生理学家 Carl Lange 从人体生理学和神经生理学出发，提出一种情绪外周理论，即把情感、情绪看作是一种心理状态和独立过程，强调自主

性内脏和神经系统对情感、情绪产生的作用。

Kleinginna 等[5]总结了近百名研究学者对情感的定义和理解，其中，两种代表性的情感定义如下。

“*Emotions constitute the primary motivational system of humans. Each of the primary emotions supplies its own unique kind of motivating information.*”

Tomkins

“*They have adaptive functions for the individual; they need to be inferred from various sources of evidence; they are based on specific cognitions; and they reveal something of an individual's attitudes and motivations.*”

Plutchnik

综上所述，研究者对于情感的定义繁复纷呈，只能在某一个有限的领域范围内取得一致。情感不仅仅是一个主观概念，同时也受到社会文化、自然环境等因素的影响，因此很难给出一个通用而准确的情感定义。然而，在很多场合下，不同人之间也确实存在着一些确定并具有一致性的情感表现信息。例如，人们在取得成功的时候，都会表现出喜悦；而在亲人去世的时候，人们都会表现出忧伤。因此，研究者的目光已经渐渐从关注繁复纷呈的情感定义方面，转向情感的表示方法方面的研究。

1.2　情感的表示方法

人类的情感是一种极其复杂的心理和生理现象，对其进行准确的定义和表示并不容易。即使在对情感的研究已经有很长历史的心理学领域，到目前为止，关于什么是情感仍没有一个十分统一的认识，也没有一个定性和定量的测量评价标准。因此，对于情感的表示，针对不同的研究目的，研究者采用不同的情感表示方法。目前，在心理学领域存在两种被广泛认可的情感表示方法，即离散的情感类别表示和连续的情感维度表示。

离散的情感类别表示是根据情感的纯度和原始度，将情感划分为基本类情感（主要情感或原始情感）和复合类情感（次要情感或派生情感）。复合类情感是由基本类情感变化混合而成的，好像三元色可以混合生成多种色彩一样，因此该情感生成理论也称为情感的调色板理论[6]。对于基本类情感到底包含哪些种类情感，至今还存有争议。这些年来，众多研究者提出的基本类情感从两种到 20 种不等。在日常生活中，通常采用日常语言标签对情感进行标识和分类，如生气、高兴、害怕等。

中国古代就将情感类型划分成 7 类，也就是常说的“七情六欲”中的七情。这七情在中国古代的典章制度书籍《礼记》中记载为“喜、怒、哀、惧、爱、恶、欲”。此外，在我国的中医基础理论中，七情则指的是“喜、怒、忧、思、悲、恐、惊”。

美国神经学科学家 Ortony 和 Tumer[7]对基本类情感类型进行了归纳和整理。不过，对于基本类情感的种类，目前比较得到认可的是 6 大类情感“the big six”[8]，即包括生气（anger）、高兴（joy/happiness）、悲伤（sadness）、害怕（fear）、惊奇（surprise）和厌恶（disgust）。在实际应用中，中性（neutral）或称为无情感也常被用到，从而构成常见的 7 种基本情感。

情感也可以用连续变化的维度表示。维度是指情感在所固有的某种性质上，存在一个可变化的度量。情感维度表示理论通常将情感定义为一个多维维度空间上的一个点。不同

研究者所定义的维度空间数目也有所不同，有二维、三维甚至四维，其中受到广泛认可的是“激发维（arousal）–效价维（valence）–控制维（power）”三维连续情感空间模型[8]，如图 1.1（a）所示。

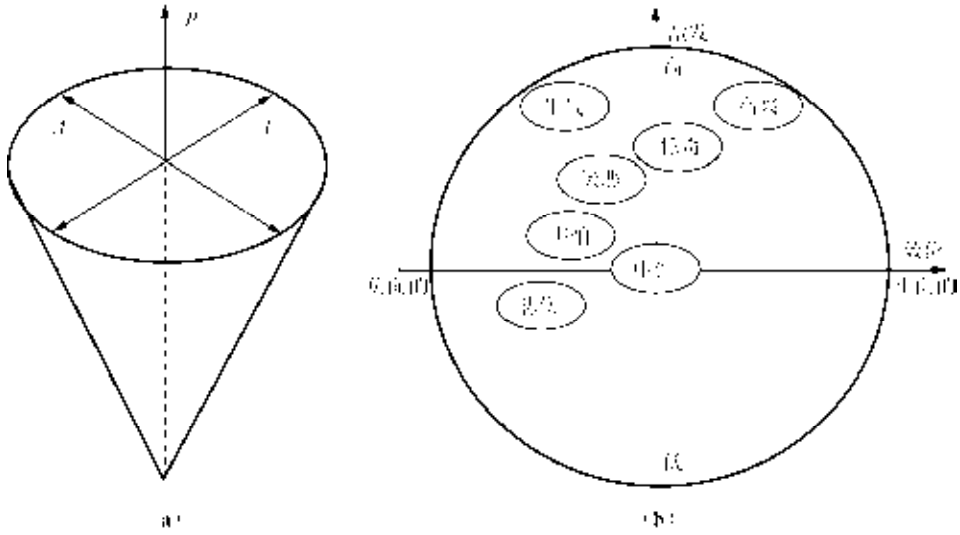

图 1.1　情感空间模型[8]

（a）三维情感空间模型；（b）二维情感空间模型

这三维情感空间模型的含义，具体表述如下。

（1）激发维。表示说话者生理上的激励程度或者对采取某种行动所做的准备、是主动的（active）还是被动的（passive）。这个维度可以表示出个体对于各种活动的参与性，是呆板的还是活跃的，是冷淡的还是兴奋的。根据这个维度的含义，可以很容易将情感状态划分等级。例如，生气和悲伤的唤醒度就明显不同，生气时主体表现出来的兴奋度比较高，因而具有较高的激发维度；而悲伤时主体的表现则不会那么兴奋，比较低调些，因此具有较低的激发维度。

（2）效价维。表示说话者对某一事物正面的（positive）或负面的（negative）评价。这个维度体现了人类情感的本质内涵。一般情况下，一个人表达情感的主要目的就是表现出其本人对他人、事或物的效价维度，即持消极还是积极的态度，或者不喜欢或喜欢的程度。效价维与情感状态之间的联系非常紧密。例如，生气作为一种消极的负面情感，因而具有较低的效价维度，而高兴是一种积极向上的情感，因而具有较高的效价维度。

（3）控制维。表示说话者的力量和控制欲望的强弱，用来区分情感状态是由主体主观上发出的，还是受客观环境影响而产生的。例如，厌恶就是主体主观上发出的，而害怕就是受客观环境的影响而产生的。

在这种三维情感空间模型中，每种情感被看成一个连续体的一部分，不同的情感被映射到这三维空间上的一个点。该点的空间坐标对应标识某一种情感。在实际应用中，研究者也经常采用简化的“激发维–效价维（arousal-valence）”二维情感空间模型如图 1.1（b）所示。Cowie 等[9]已经开发了一种 Feeltrace 工具，用于实时测量视频片段中情感内容的“激发维–效价维”二维情感空间坐标。

尽管这种情感维度表示理论未必能够全面反映情感的各个侧面，但是提供了一种不同

情感状态之间“距离”的简单度量方法，因而可以对情感进行某种程度的分类。此外，这种情感维度表示理论也提供了一种情感的连续表达方法，可以方便地实现对语音情感的连续性变化的跟踪。因此，这种情感维度表示理论不仅在心理学领域产生了较大的影响，而且在情感信息处理领域也备受研究者的关注。

Pereira[10]通过 31 人的听辨实验研究确定了 happiness、sadness、cold anger、hot anger 及 neutrality 这 5 种情感状态的激发维度、效价维度和控制维度 3 个空间维度上的等级，并发现情感维度的概念非常有利于描述和区分不同情感的类别。此外，研究也表明激发维度上比较接近的情感，如 anger 和 happiness，是最容易混淆的。Tato 等[11]通过研究语音特征参数（如韵律特征和音质特征）与二维情感空间（arousal-valence）的关系，用于改善情感识别性能。Wu 等[12]研究了基频、能量、发音持续时间和 Mel 频率倒谱系数（mel-frequency cepstral coefficients，MFCC）等声学特征参数与三维情感空间之间的关系，发现 MFCC 参数与三维情感空间的关联度最高，其次是能量、基频和发音持续时间。

1.3　情 感 计 算

1985 年，人工智能创始人之一，美国麻省理工学院 Minsky 教授在其《脑智社会》（The Society of Mind）专著[13]中指出，“问题不在于智能机器能否有情感，而在于没有情感的机器能否实现智能”（The question is not whether intelligent machines can have any emotions, but whether machines can be intelligent without emotions）。1995 年，美国神经生理学家 Damasio[14]在对大脑神经活动的研究中发现，人类的智能不仅表现为正常的理性思维和逻辑推理能力，也应表现为正常的情感能力。情感能力是人类智能的重要标志，情感在人与人之间的交流中必不可少。实际上，中国古代的老话“晓之以理，动之以情”“合情合理”“尽在情理之中”也都表达了同样的意思，即人们在决策、处事、待人时，“情”和“理”常常是密不可分的。

传统的人机交互，主要通过键盘、鼠标、屏幕等工具进行，只追求便利和准确，无法理解和适应人的情绪或心境。而如果计算机缺乏这种情感理解和表达能力，就很难指望计算机具有类似人一样的智能，也很难期望人机交互做到真正的和谐与自然。由于人类之间的沟通与交流是自然而富有感情的，因此在人机交互过程中人们也很自然地期望计算机具有情感能力。

1997 年，美国麻省理工学院 Picard 教授提出了“情感计算”[15]（affective computing）的概念，即将“情感计算”定义为“关于、产生于或故意影响情感方面的计算（computing that relates to, arises from, or deliberately influences emotions）”。情感计算研究的目的就是要赋予计算机类似于人的观察、理解和生成各种情感特征的能力，最终使计算机像人一样能进行自然、亲切和生动的交互。

目前，关于情感计算的研究是一门综合认知科学、生理学、心理学、语言学、计算机科学等多学科交叉的热点研究课题，正越来越受到国内外科研机构和研究人员的重视。1997 年，美国麻省理工学院媒体实验室建立了世界上第一个情感计算小组（https://affect.media.mit.edu/），专门侧重于有关情感信号的获取（如各类传感器的研制）与识别。IBM 公司研究开发“情感鼠标”，可根据手部的血压及温度等传感器感知用户的情感，并作出适当的反应。美国卡内基 • 梅隆大学专门研究开发基于情感计算的可穿戴式计算机，致力于研

究情感计算技术的实际应用。为推动我国在此国际前沿方向的研究，2004 年国家自然科学基金委将情感计算理论与方法的研究首次列入当年拟资助的重点项目指南中。2005 年，中国科学院自动化研究所、中国自动化学会、中国计算机学会、中国图象图形学会、中国中文信息学会、国家自然科学基金委员会以及国家“863 计划”计算机软硬件技术主题等部门作为主办单位在北京主办了首届国际情感计算及智能交互（affective computing and intelligent interaction, ACII）学术会议。2009 年，由清华大学心理学系、中国科学院心理学研究所等部门在南昌联合主办的首届全国认知科学研讨会，首次将“情感计算”列为认知科学领域当前重点关注的前沿课题之一。2010 年，为了促进“情感计算”这个新兴主题的研究和发展，IEEE 计算机协会新创办了一个以“情感计算”命名的国际学术期刊 *IEEE Transactions on Affective Computing*。2015 年，备受关注的日本软银公司开发的情感机器人“Pepper”和微软研究院发布的第三代聊天机器人“小冰”3.0，由于都采用了情感计算技术，如通过语音和人脸表情来分析出用户的情绪，从而帮助机器人与用户之间的交互更加类人化。2017 年 7 月，在《国务院关于印发新一代人工智能发展规划的通知》中，感知智能是十分核心的研究领域，而情感计算则是感知智能技术中非常重要的组成部分。可见，有关情感计算的研究是当前人机交互、计算机科学、认知科学等领域亟待开垦的沃土之一。

情感计算的实现需要一个完整的情感交流过程。这就涉及一系列相关的理论和技术，其核心技术可以分为情感信号的获取、情感识别、情感理解与反馈以及情感表达 4 大部分[16]。

情感信号的获取主要研究各类有效传感器设备的研制，以便记录下情感信号。情感信号主要包括语音、面部表情、站姿、手势等体态语，以及皮肤电阻、脉搏等生理指标。美国麻省理工学院媒体实验室在传感器方面的研制走在了世界前沿，他们已经研制出皮肤电流传感器、脉压传感器、肌电流传感器、汗液传感器等多种传感器。皮肤电流传感器可以用来实时测量皮肤的电导率，然后根据电导率的变化就可以测量用户的紧张程度。脉压传感器可以用来实时监测由心动变化而引起的脉压变化。肌电流传感器用来测量肌肉运动时的弱电压值。汗液传感器通过其本身伸缩的变化来测量呼吸与汗液的关系。IBM 研制的情感鼠标可测量出人体的脉搏、体温、皮肤电反应等生理指标。

情感识别是指对采集的情感信号进行分析建模，识别出人类内在的情感类型。情感识别是情感计算研究中最基础、最重要的内容之一。情感作为一种内部的主观体验，总是伴随着某种外部表现特征，即表情。借助表情，人们才能“察言观色”，在别人的举手投足间洞悉他人的内心感受。根据表情的发生部位和方式的不同，人的表情主要分为 3 种，即言语表情、面部表情及身体姿态表情。言语表情是指利用说话人的声调、节奏、速度等变化来表示情感，因此言语表情又叫作“情感语音”。面部表情是指采用人脸的面部肌肉的变化来表示情感，因此面部表情又叫作“人脸表情”。其中，对于言语表情的识别，一般称为“语音情感识别”或“情感语音识别”；对于面部表情的识别，一般称为“人脸表情识别”或“面部表情识别”；对于身体姿态表情的识别，一般称为“姿态表情识别”。由于人类的身体姿态表情（如运动姿势）变化方面的规律性，比言语表情和面部表情更难获取，所以目前情感识别的研究重点是语音情感识别和人脸表情识别。同时融合言语表情和面部表情的情感识别，通常称为“音视频情感识别”或“听视觉情感识别”。而融合两种及以上的表情识别，通常称为“多模态情感识别”。因此，音视频情感识别是一种相对比较简单的多模态情感识别类型。

情感理解与反馈是指计算机通过分析用户情感产生的原因，对用户的情感变化作出最适宜的反应。例如，当用户厌烦时，计算机就会积极主动提供更加新鲜而有趣的内容。当用户困惑时，计算机就会及时地鼓励用户，并且提供对用户很有价值的建议。情感理解与反馈是情感计算技术研究的核心，只有计算机具备了这样的能力，才能真正实现人机交互的自然化。

情感表达是指给予某一种情感状态，研究如何使这一情感状态在一种或几种行为或生理特征中体现出来。也就是要求计算机交互设备能够向用户模拟和表达出情感。对拟人化的情感表达的研究是目前的主流，如通过动画、语音等方式表达情感。现在，计算机可以综合运用语音、面部表情、肢体语言等模态信息，表达出生气、高兴、悲伤等情感，甚至能在一定的情境下表达出困惑、同情等细微情感。在人脸面部图像合成领域，现代的计算机已经能够合成出具有混合表情特征的面部图像。同样，由计算机合成的语音也不再那么机械和单调，而是能够体现出生气、高兴、悲伤等不同的情感语音基调。

1.4　情感识别的应用

情感识别是情感计算领域中最基础、最重要的研究内容之一。根据情感的外部表现特征方面的“表情”，情感识别的研究主要包括 3 个方面，即基于言语表情的语音情感识别、基于面部表情的人脸表情识别以及融合多种表情信息的多模态情感识别，如音视频情感识别。

情感识别的研究具有重要的应用前景，其研究的最终目标是让计算机通过情感信号对用户的情感信息进行获取、识别和响应，以便帮助用户在和谐、自然的人机交互模式下高效地完成既定的任务。该研究具体的应用可以分为以下 4 类。

（1）在服务业，应用情感识别技术的自动远程电话服务中心通过理解客户的“画外音”及时发现客户的不满情绪，使得公司能够及时有效地做出变通，最大限度地保留住可能的客户资源。

（2）在教育领域，具备了情感识别能力的计算机远程教学系统，及时识别学生的情绪并做出适当的处理，从而提高教学质量。

（3）在医学领域，拥有情感识别功能的机器人能帮助那些缺乏正常情感反应和交流的孤僻症患者反复练习情感交流，逐步达到康复的目的。情感识别技术用于机器人领域，将赋予机器人具有表达、识别和理解人类的喜怒哀乐，模仿、延伸和扩展人的情感能力。

（4）在娱乐业，可用于视频点播系统、电子宠物和游戏动漫。例如，结合情感识别技术的视频点播系统能对广播电视节目进行情感标注，根据用户提交的情感需求做出合理的响应，使得用户能随心所欲地看到“高兴”或“难过”的节目。拥有双向情感交流能力的电子宠物将类似于一个真实的动物宠物。它能丰富人们的生活，帮助孩子学习与生物的情感交流。在计算机游戏动漫系统中加入情感识别交互技术，则能够构筑更加拟人化的风格和更加逼真的虚拟场景。这样一方面可以降低玩家的疲劳度，另一方面又能给予玩家更全面的感官享受，增加游戏的娱乐性。

第 2 章　基于语音和人脸的音视频情感识别技术回顾

面向语音和人脸的音视频情感识别是“情感计算”领域中情感识别研究方向的一个重要分支，近年来发展比较快，前景广阔。本章回顾了该方向涉及的单一模态的语音情感识别、人脸表情识别、融合语音和人脸的音视频情感识别等方面的研究重点，如情感数据库、情感特征分析以及情感的分类算法等，也简要介绍了近年来发展起来的深度学习理论在音视频情感识别中的应用现状。

2.1　语音情感识别的技术回顾

言语表情是通过语音的高低、强弱、抑扬顿挫等变化来表达说话人的情感的。在人际交往中，语音作为语言的声音表现形式，是人类交流最自然、最方便、最有效的手段之一。人类的语言不仅包含了文字符号信息，同时也携带着人们的感情和情绪等信息。通过语音的高低、强弱、抑扬顿挫等变化，人们不仅很容易表达出自己的情感变化，也同样很容易感受到对方的情感变化。说同样一句话，往往由于说话人的情感状态不同，其意思和给听者的感觉就会不同。譬如，“你真行”这句话，发音时运用不同语气，可以使之成为一句赞赏的话，也可以使之成为妒忌或讽刺的话。因此，如何让计算机通过语音信号自动分析和判断说话人的情感状态，即“语音情感识别”方面的研究就显得尤为重要。

当前，语音情感识别研究的重点主要集中在情感语音数据库的建设、情感声学特征的提取和分析、语音情感分类方法等方面[17, 18]。国外对语音情感识别的研究起步于 1990 年，研究最活跃的是美国麻省理工学院媒体实验室的情感计算小组。1990 年，美国麻省理工学院媒体实验室情感计算小组的 Cahn[19]开发了世界上第一个情感语音编辑器“Effect Editor”，首次描述了基频、语速、音质特征等声学参数与情感状态的关系，为语音情感识别研究的声学特征分析打下了基础。国内对于汉语普通话的语音情感识别的研究起步于 2000 年之后，具代表性的研究单位有中国科学院自动化研究所、清华大学、哈尔滨工业大学、浙江大学以及东南大学等。

在语音情感识别研究方面，具代表性的研究工作者主要包括美国南加利福尼亚大学的 Narayanan[20]、德国帕绍大学的 Schuller[21]、中国科学院自动化研究所的陶建华[22]、清华大学的蔡莲红[23]、东南大学的赵力和郑文明[24, 25]、哈尔滨工业大学的李海峰和韩纪庆[26]、华南理工大学的文贵华[27]、江苏大学的毛启容和詹永照[28]等。

2.1.1　情感语音数据库

情感语音数据库是进行语音情感识别研究的基础，而且情感语音数据库的好坏直接影响到最后的情感识别效果的可靠性。一个高质量的情感语音数据库的建立[29]必须符合以下 4 个条件。

（1）真实性（genuine）：情感素材应能够反映人们真实的情感感受。

（2）丰富性（richness）：情感素材应包含载有情感信息的语音、面部表情等多媒体信息。

（3）交互性（interaction）：情感素材应取之于人与人之间交互过程中产生的样本。

（4）层次性（gradation）：情感素材应包含以日常生活中的方式产生的各种典型情感。

根据情感素材的情感自然度程度的不同，目前研究者们建立情感语音数据库所用的方法可分为以下 3 种。

（1）自然语音（spontaneous speech）：从现实生活中采集真实的自然情感语料，通过人工筛选获得可用的语料。

（2）模拟语音（acted speech）：让专业或非专业人士通过情感模仿进行语料的录制。

（3）诱导语音（elicited speech）：营造恰当的环境氛围刺激专业或非专业人士，然后进行语料的录制。

这 3 种情感语音各有优、缺点。

（1）自然语音来自于现实生活，是人们在现实生活中表现出最真实情感的语音。但是，采用这种方法获得自然情感语音数据库非常困难，因为理想的情况下要求参加录音的人必须不知道自己正在被录音，这就涉及很多的社会伦理道德问题。目前还没有存在这种理想的自然情感语音数据库的报道，只是研究者普遍认为这是最真实的情感语音来源。一种建立这种自然情感语音数据库的可替代方法是从一些电视、电影等媒体材料收集一些非常自然的情感对话片段。但用这种方法获得自然情感语音数据库的工作量比较大，而且大部分媒体材料除了语音之外，对话还经常伴有背景音乐等杂音，因此适合试验要求的素材比较难找。

（2）模拟语音是最为常见的一种情感语料。大多数研究者采用了人工模拟情感的语料进行语音情感识别方面的研究。这种模拟情感语料具有两个显著优点：①可操作性很强，只需要一些简单的录音设备就可以在一个安静的录音环境里快速地完成所需的语料录制任务；②这种录制的语料符合性别要求、文字要求和情感要求，而且情感可区分性也较好。尽管使用这种简易录制的模拟情感语音数据库通常能够获得较高的语音情感识别性能，但这种模拟情感语料中的情感成分往往被夸大，其情感的自然度与现实生活中的真实情感还具有较大差距，并不能真正地体现出人类在自然环境中的真实情感。

（3）诱导语音介于上述两种语音之间，其可操作性也比较强，并且可以获得比使用第二种方法更为接近真实情感的情感语料，但是这种方法无法确认环境对录音者的刺激是否有效以及刺激所起的作用有多大。

目前，国外研究者已经建立了包含各种语言的离散或维度情感类型的情感语音数据库，如英语、德语、波兰语、日语、荷兰语、西班牙语、丹麦语、瑞典语、俄罗斯语等，详细情况可见综述类文献[30-32]。现有的国外情感语音数据库数量已达数十个，其中具有代表性的离散情感语音数据库如表 2.1 所示。有关代表性的维度情感语音数据库，主要有 VAM[33]（德语）、IEMOCAP[34]（英语）及 RECOLA[35]（法语）数据集。

表 2.1　具有代表性的国外情感语音数据库（离散情感）

数据库名称	语言	自然度	情感类型	人数	语句数量	多媒体信息
Berlin Emotional Speech Database (EMO-DB)[36]	德语	模拟	生气、烦躁、厌恶、害怕、高兴、悲伤、中性	女性 5 人 男性 5 人	535	音频
Danish Emotional Speech (DES)[37]	丹麦语	模拟	生气、高兴、中性、悲伤、惊奇	女性 2 人 男性 2 人	419	音频

续表

数据库名称	语言	自然度	情感类型	人数	语句数量	多媒体信息
Speech Under Simulated and Actual Stress (SUSAS)[38]	英语	模拟和诱导	害怕、中性	女性 3 人 男性 4 人	1185	音频
eNTERFACE05[39]	英语	模拟	生气、厌恶、害怕、高兴、悲伤、惊奇	女性 8 人 男性 34 人	1277	音视频
RML[40]	6 种语言	模拟	生气、厌恶、害怕、高兴、悲伤、惊奇	男性 8 人	720	音视频
SmartKom[41]	德语和英语	自然	生气、高兴、中性、无助、沉思、惊奇	女性 47 人 男性 32 人	2775	音视频
Audio-visual Interest Corpus (AVIC)[42]	英语	自然	中性、烦躁、高兴	女性 10 人 男性 11 人	996	音视频
BAUM-1s[43]	土耳其语	自然	生气、厌恶、害怕、高兴、悲伤、惊奇	女性 17 人 男性 14 人	521	音视频
AFEW5.0[44]	英语	自然	生气、厌恶、害怕、高兴、悲伤、惊奇、中性	330 人	1645	音视频

近年来，国内研究者也相继建立了一些汉语普通话情感语音数据库。这些建立的汉语情感语音数据库，大都以模拟语音方式录制所需的情感语料，如中国科学院自动化研究所的陶建华[22]、清华大学的蔡莲红[23]、浙江大学的陈纯[45]、东南大学的赵力[24]、哈尔滨工业大学的韩纪庆[26]等。此外，赵力等[46]通过计算机游戏诱发的方式，建立了一个情感自然度较高的诱导语音情感数据库。近年来，李雅和陶建华等[47]通过收集一些电视访谈、电影等片段材料建立了一个情感自然度非常高的自然情感音视频数据库。表 2.2 列出了上述国内研究者建立的汉语情感语音数据库情况。

表 2.2　具有代表性的汉语情感语音数据库（离散情感）

研究者	自然度	情感类型	人数	语句数量	多媒体信息
陶建华[22]	模拟	生气、害怕、高兴、悲伤、惊奇、中性	女性 2 人 男性 2 人	7200	音频
蔡莲红[23]	模拟	生气、害怕、高兴、悲伤、惊奇、中性	女性 1 人	1200	音频
陈纯[45]	模拟	生气、高兴、悲伤、惊奇、中性	女性 2 人 男性 2 人	1500	音频
韩纪庆[26]	模拟	生气、高兴、悲伤、惊奇	女性 9 人 男性 5 人	1256	音频
赵力[24]	模拟	生气、高兴、惊奇、悲伤	男性 10 人	3000	音频
赵力[46]	诱导	烦躁、喜悦、平静		1200	音频
李雅与陶建华等[47]	自然	生气、高兴、担心、急切、厌恶、悲伤、惊奇、中性	女性 113 人 男性 125 人	2852	音视频

2.1.2　语音情感特征分析

语音情感特征分析主要包括声学特征提取和声学特征降维。采用何种有效的语音情感特征参数用于情感识别，是语音情感识别研究最关键的问题之一，因为所用的情感特征参数的优劣直接决定了最终的情感识别结果的好坏。

1. *声学特征提取*

尽管目前对于到底哪些声学特征参数是最重要的语音情感特征参数，在语音情感识别领域还存在争论，但常提取的语音情感声学特征参数主要有 3 种，即韵律特征、音质特征及谱特征。

根据心理学和语音学的研究，语音信号中的情感特征一般是通过与发音语调、轻重相关的韵律表现出来的。例如，当一个人很生气的时候，说话的速度会变快，音量会变大，音调会变高。相反，当一个人很悲伤的时候，讲话的语速会变慢，音量会变小，音调会变低。这些韵律的变化都可以很直观地被人们所感受到。因此，在早期的语音情感识别研究文献[24, 48]中，首选的声学特征参数是韵律特征，如基音频率（基频）、振幅（能量）、发音持续时间、语速等。实际使用韵律特征时，一般采用的是在这几种特征的基础上所衍生的一些统计学参数，如均值、标准差、范围、中值等。这些韵律特征能够体现说话人的部分情感信息，较大程度上能区分不同的情感类型。因此，韵律特征已成为当前语音情感识别中使用最广泛并且必不可少的一种声学特征参数[49]。

除了韵律特征，另外一种常用的声学特征参数是与发音方式相关的音质特征参数。Tato等[11]从语音特征与三维情感空间模型之间的关系出发，研究指出若提取的情感特征信息反映的情感空间维数越多，就越能更好地区分不同的情感。韵律特征主要表达三维情感空间模型中的“激发维”信息，因而使用韵律特征可以较好地区分在“激发维”上差异明显的情感类型，如中性和惊奇。但对于在“激发维”上比较接近的情感类型，如生气和高兴，仅使用韵律特征来识别是不够的。文献[11]的研究表明，语音信号中的音质特征，如共振峰、频谱能量分布、谐波噪声比等，不仅能够很好地表达三维中的“效价维”信息，而且也能够部分反映三维中的“控制维”信息。因此，为了更好地识别情感，同时提取韵律特征和音质特征，尤其是在这方面研究提取一些新的特征参数用于语音情感识别，已成为语音情感识别领域声学特征提取的一个重要研究方向[50, 51]。笔者在这方面也作了一些探讨，文献[52]发现将提取的音质特征参数和韵律特征参数相结合所取得的语音情感识别性能，比单独使用韵律特征参数时高出 6%。近年来，声门特征（glottal features）[53]和声源特征（voice source features）[54]也被用作新的音质特征参数用于情感识别，并表现出良好的性能。

谱特征参数是一种能够反映语音信号的短时功率谱特性的声学特征参数，如 LPC（linear prediction coefficients）、LPCC（linear prediction cepstral coefficients）、LFPC（log frequency power coefficients）、MFCC（mel-frequency cepstral coefficients）等。其中，MFCC 是最具代表性的谱特征参数，被广泛应用于语音情感识别[55-58]。由于谱特征参数及其导数，仅反映语音信号的短时特性，忽略了对情感识别有用的语音信号的全局动态信息。近年来，为了克服谱特征参数的这种不足，研究者提出了一些改进的谱特征参数，如类层次的谱特征（class-level spectral features）[56]、调制的谱特征（modulation spectral features）[59]和基于共振峰位置的加权谱特征（weighted spectral features）[60]等。

对于常见的声学特征参数，如韵律特征、音质特征及谱特征，与不同情感之间的具体关系，研究者也做了不少工作。Murray 和 Arnott[61]较早总结了生气、高兴、悲伤、害怕和厌恶 5 种基本情感，与韵律特征、音质特征变化之间的关系，如表 2.3 所示。

表 2.3　声学特征与不同情感的关系[61]

情感 / 特征	生气	高兴	悲伤	害怕	厌恶
speech rate	slightly faster	faster or slower	slightly slower	much faster	very much faster
pitch average	very much higher	much higher	slightly lower	very much higher	very much lower
pitch range	much wider	much wider	slightly narrower	much wider	slightly wider
intensity	higher	higher	lower	normal	lower
voice quality	breathy, chest	breathy, blaring tone	resonant	irregular voicing	grumble chest tone
pitch changes	abrupt on stressed	smooth, upward inflections	downward inflections	normal	wide, downward terminal inflects
articulation	tense	normal	slurring	precise	normal

由表 2.3 可知，生气、高兴、悲伤、害怕和厌恶 5 种基本情感与声学特征参数之间的关系如下。

（1）生气。当人生气时，其生理特征比平时要突出，如心跳加快、血压升高、皮肤电压升高等。这同时会对声学特征参数的变化产生影响。由于生理特征发生变化，人生气时胸腔的呼吸声和回声就会在语音信号中所占的比例有所增加，从而导致语音的振幅强度比普通的情感要高得多，语速也比普通的语句要快。为了突出生气的效果，基音就会在重音处发生语调的突变，这就变成人处于生气状态的一个重要特征。

（2）高兴。当人高兴时，发音时的语音信号中的语速一般难以精确确定，在不同的情况下会有不同的表现特征。高兴与生气的生理特征的共同之处在于，它的声音中也常带有呼吸声。与其他情感的不同之处主要在于，人高兴时的基音变化曲线通常是向上弯曲的。高兴时的语音振幅强度往往集中在句子末尾的一两个字上，整个语句声调的调阈较平静语句要高。高兴语句的前部与中部要比相应内容的平静语句的语速稍微快一些。

（3）悲伤。当人悲伤时，一般都比较压抑，因此带悲伤的语句的时长比平静语句要慢，其振幅强度也比其他各种情感低得多。人悲伤时的基音变化曲线是向下弯曲的。由于悲伤时的语速比较慢，字与字之间的读音彼此都隔得比较开，导致字调的调形保留了其原来单字的调形，从而使得多字调的效果被弱化。此外，在悲伤语句中，几乎所有的字中都夹杂着多多少少的鼻音，这就需要作去鼻音化处理，使得悲伤语句的调阈降低，整个语句也趋于平坦化。

（4）害怕。当人害怕时，语句的语速、基音、基音范围等特征，和生气、高兴两种情感的语句相类似，不同之处在于害怕语句的清晰度要比其他情感精确。

（5）厌恶。由于厌恶和生气具有非常高的相似性，这两种情感语句的语音特征参数比较接近，因此有些研究者对厌恶和生气这两种情感并不作区分，而是把厌恶归入到生气类情感中进行研究。厌恶和生气的主要区别在于，厌恶语句的基音变化率较宽，并在其语句的末端具有向下倾斜的变化趋势。

Cowie 等[48]则进一步总结了 14 种不同的情感类型与韵律特征、音质特征和谱特征的关系。赵力等[24]也研究了喜、怒、惊、悲 4 种情感的时间构造、振幅构造、基频构造、共振峰构造的分布规律。韦岗等[62]也总结了声学特征参数与情感之间的关系。Luengo 等[63]对韵律特征、音质特征以及谱特征在语音情感识别中的重要性进行了全面分析，发现谱特征的作用高于韵律特征、音质特征。

上述提到的声学特征基本上都是线性特征，已有一些非线性特征受到研究者的关注，其中最具代表性的是非线性 Teager 能量算子（teager energy operator，TEO）。

根据语音信号生成的物理模型，声道被认为是一根不均匀的管子。线性的语音分析理论是假设沿管轴方向传播平面波的，而基于 TEO 的非线性语音分析理论认为，发音时的气流通过声带和伪声带区域就会出现气流的分离与附着，从而形成涡流，并与平面波一起构成语音信号生成的主要原因，并且是非线性的。根据上述非线性语音分析理论，Teager[64]通过实验研究提出了一种能够反映这种涡流的非线性作用的 Teager 能量算子，即称为 TEO。

对于离散的时间信号，TEO 被定义为[64]

$$\psi\left[x(n)\right] = x^2(n) - x(n+1)x(n-1) \tag{2.1}$$

由式（2.1）可知，信号 $x(n)$ 在 n 点的 TEO 只与该样本点以及其前后各一个样本点有关系，其运算也非常简单。对于含有情感的语音信号，根据上述非线性理论，声激励源的变化必然包括线性和非线性两种成分。因此，将非线性 TEO 引入到特征参数，理论上是可以从非线性和线性两个角度对不同感情影响下的语音变化进行分析研究。鉴于上述非线性 TEO 的良好特性，研究者已成功将 TEO 应用到语音情感识别中，并取得了良好的性能[65, 66]。

2. 声学特征降维

特征降维是指通过映射或变换方式将高维特征空间映射到低维特征空间，从而达到降维的目的。特征降维算法分为线性和非线性两种。最具代表性的两种线性降维算法，如主成分分析（principal component analysis，PCA）和线性判别分析（linear discriminant analysis，LDA），已经被广泛用于对语音情感特征参数的线性降维处理[55, 67]。

PCA 是由 Jolliffe[68]提出的，是一种根据已知数据的方差而进行最优降维表示的方法。该方法的主要思想是根据样本数据点在多维模式空间不同的位置分布情况，用样本数据点在多维空间中变化的最大方向（方差最大的方向）作为判断数据压缩好坏的标准，并实现最优数据子集的特征提取。

LDA 最初是由 Fisher[69]提出的，其主要思想是通过最大化样本数据的类间散度和最小化样本数据的类内散度来选取投影方向，从而达到数据降维的目的。

近年来，已有的语音学研究[70-72]指出，从人的发音机理来看，由多个不同截面积的管子串联而成的声道（vocal tract）系统产生的语音信号刚好位于一个嵌入在高维声学特征空间的低维非线性流形上。他们发现具有内在的低维非线性流形结构的语音信号能够很好地反映人的发音本质。因此，情感语音的特征数据应该也具有内在的低维非线性流形结构。考虑到语音情感特征数据这种非线性流形结构的特点，仍然采用线性特征提取方法（如 PCA 和 LDA）来处理语音情感特征数据可能就不太适用了。

为了解决上述问题，流形学习（manifold learning）[73]方法提供了可能的解决方案。它

试图将人类的认知流形规律引入到机器学习领域中，从而使机器能够学习出高维数据空间的几何结构或内部规律。流形学习的目的就是要研究和模拟人类的这种感知能力，以从有限的离散样本数据中发现嵌入在高维空间中的低维非线性流形结构，并给出一个有效的低维表示。两种代表性的流形学习方法是局部线性嵌入（locally linear embedding, LLE）[74]和等距映射（isometric mapping, Isomap）[75]。目前，这两种流形学习算法都被应用于语音情感特征数据的非线性降维处理[71, 76]。但这些原始的流形学习方法直接应用于语音情感识别中的特征降维，所取得的性能并不令人满意。主要原因是它们都属于非监督式学习方法，没有考虑对分类有帮助的已知样本数据的类别信息。

为了增强流形学习方法的学习能力，发展新的流形学习算法，用于改善语音情感识别的性能近年来备受关注。You 等[22]提出一种增强型 Lipschitz 嵌入（enhanced Lipschitz embedding，ELE）算法，在特征降维方面的性能优于 LLE 和 Isomap。陈纯等[45]提出了一种基于协方差描述子和黎曼流形的非线性特征降维方法用于语音情感识别。此外，发展具有判别学习能力的监督流形学习方法，进一步提高流形学习方法所产生的低维嵌入数据的判别力，从而改善语音情感识别性能，也是一个值得研究的方向。近年来，我们也在流形学习方法及其在语音情感识别应用方面[77-82]进行了一些探讨，其中具有代表性的工作详见第 3 章和第 4 章。

2.1.3　语音情感的分类算法

语音情感识别本质上是一个模式识别问题，所以几乎所有的模式识别方法都可以应用于语音情感识别。20 世纪 90 年代，率先应用于语音情感识别的方法是线性判别分类器（linear discriminant classifier，LDC）[83, 84]和 *K* 近邻法（K-nearest neighbor，KNN）[84]。到 2000 年，人工神经网络（artificial neural network，ANN）[85]开始成为一种流行的语音情感识别方法。到 2002 年，具代表性的语音情感识别方法主要有支持向量机（support vector machines，SVM）[86]、隐马尔可夫模型（hidden Markov models，HMM）[87]和高斯混合模型（Gaussian mixture models，GMM）[88]。每一种模式分类器都有自己的优、缺点。融合多个分类器优点的分类器组合方法[89, 90]也受到研究者的关注。此外，近年来，我们也尝试利用压缩感知的稀疏表示思想，探讨了稀疏表示分类方法、核稀疏表示分类方法在语音情感识别中的应用[91, 92]。

下面通过实例分别叙述 LDC、KNN、ANN、SVM、HMM 和 GMM 在语音情感识别中的应用。

1. 线性判别分类器

线性判别分类器（LDC）是一种基于线性判别函数的模式分类器。这种分类器计算简单，不要求估计特征向量的类条件概率密度，是一种非参数分类方法。Banse 等[83]选用了 12 名表演者模拟发音，建立了一个包含生气、高兴、悲伤等 14 种情感类型的 1344 句大小的德语数据库，提取基频、能量、语速以及有声与无声的长时平均频谱等声学参数，采用 LDC 分类器取得了 50%左右的正确识别率。Lee 等[55]将基频、能量、语速、第一共振峰、第二共振峰等声学信息与词汇、语义信息相结合，识别正面（如高兴、中性）和负面（如生气、悲伤）两大类情感，对 7200 句的电话呼叫中心的英语情感对话进行情感识别测试，采用 LDC 分类器取得的正确识别率接近 90%。

2. *K* 近邻法

K 近邻法（KNN）是一种基于样本学习的非参数化模式分类器，分类时直接从训练样本中找出与测试样本最接近的 *K* 个样本，以判断测试样本的类属。该分类器计算简单，容易实现。Dellaert 等[84]提取基频、语速等参数，在包含生气、悲伤、高兴和害怕 4 种情感的 1000 句大小的英语情感语音数据库中，比较了最大似然贝叶斯分类法（maximum likelihood Bayes classification）、核回归（kernel regression)和 KNN 等 3 种分类器的识别性能，结果发现 KNN 表现最好，能取得 70%左右的正确识别率。为了进一步提高 KNN 的分类性能，Attabi 和 Dumouchel[93]提出了一种改进的 KNN 算法（weighted ordered classes-nearest neighbors，WOCNN）用于语音情感识别，并表现出优越的性能。

3. 人工神经网络

人工神经网络（ANN）是由大量相连的神经元构成的大规模并行计算系统，通过训练过程来学习复杂的非线性输入输出关系。经典的 ANN 主要有 3 种，即多层感知器（multi-layer perceptron，MLP）、循环神经网络（recurrent neutral network，RNN）和径向基神经网络（radial basis function neutral network，RBFNN）。其中，与 MLP 和 RNN 相比，RBFNN 计算相对简单。目前，这 3 种神经网络都已经被应用于语音情感识别。

Nicholson 等[85]采用一种 one-class-in-one（ONOC）的网络拓扑结构，即为每一种情感训练一个子网络，每个子网络是一个多层感知器，提取基频、线性预测系数（LPC）及导数作为特征参数输入到每个子网络中，再根据各个子网络的输出进行决策得到最终的情感识别结果。在包含 100 人的 640 000 语句的日语情感数据库中识别 8 种情感（高兴、戏弄、害怕、悲伤、厌恶、生气、惊奇、中性）取得了 50%左右的正确识别率。Park 等[94]提取基频特征，输入到一个具有一个输入节点、两个隐层节点和 4 个输出节点的 RNN 网络，在一个包含 4 种情感（中性、生气、嘲笑、惊奇）且只有 22 句语音的小样本数据库的情感识别测试中取得的最高正确识别率接近 83%。Zhang 等[95]将其与 LDC、KNN、C4.5 决策树相比，使用韵律特征和音质特征，在包含生气、高兴、悲伤和中性的 800 句的汉语普通话情感语音数据库的测试中，发现 RBFNN 好于其他 3 种方法。

4. 支持向量机

支持向量机（SVM）是一种基于统计学习理论的机器学习方法，其基本思想是将原始的数据空间通过一个核函数映射到一个高维特征新空间，从而在这新的空间构建最优分类超平面实现数据的最优分类。由于 SVM 是在结构风险最小化原则上建立起来的，从而保证其学习具有良好的泛化能力，即使对小样本训练数据也可以得到较好的性能。近年来，SVM 已经被广泛应用于语音情感识别中，成为一种有效的语音情感识别分类器。

Luengo 等[63]在 535 句模拟的 Berlin 德语情感语音数据库上识别生气、高兴、悲伤、惊奇、害怕、厌恶及中性 7 种情感，提取基频、能量、语速、频谱、共振峰等声学特征，利用 SVM 取得了 78%左右的正确识别率。Sobol-Shikler 等[96]提取基频、能量、语速和谐波属性等声学特征，在包含有趣、兴奋、自信等 9 种复杂情感的 633 句模拟的英语情感语音数据库进行情感识别测试，采用 SVM 取得的正确识别率为 83%。Schuller 等[67]结合声学特征和语言信息，比较了 LDC、KNN、MLP、GMM 和 SVM 的性能，发现 SVM 优于其他方法。

值得指出的是，SVM 模型中的核函数直接影响着 SVM 的分类性能。如何根据训练样本数据选择和构造合适的核函数，以及确定核函数的最佳参数等问题，至今仍然缺乏相应的理论指导。实际应用中，大多采用交互式验证方法在训练样本数据集上进行搜索核函数的最佳参数，如常用的径向基核函数的核宽度参数，这会导致训练时间较长。

5. 隐马尔可夫模型

隐马尔可夫模型（HMM）是一种基于转移概率和传输概率的随机模型。由于 HMM 能够很好地描述语音信号的整体非平稳性和局部平稳性，现已被广泛应用于基于时序特征的语音情感识别模型中。按照 HMM 的状态转移概率矩阵，HMM 可分为遍历型和从左到右型。一般而言，遍历型 HMM 适合于与文本无关的语音情感识别，而从左到右型 HMM 适合于与文本有关的语音情感识别。按照 HMM 的输出概率分布，HMM 可分为离散型、连续型和半连续型。离散型 HMM 模型简单，计算量较少，但必须对语音情感特征参数进行矢量量化处理。这就容易造成部分信息的丢失，从而影响系统的情感识别精度。连续型 HMM 可以直接处理语音情感特征参数，不需要矢量量化，但使用时需要较多的概率密度函数和训练数据样本，从而导致模型复杂，运算量大，训练时间较长。半连续型 HMM 的特点介于离散型和连续型之间。

Nwe 等[87]使用 4 个状态的遍历离散型 HMM，提取 LFPC 特征参数，在包含 12 人的 720 句汉语普通话情感语音数据库识别 6 种情感（生气、厌恶、害怕、高兴、悲伤、惊奇）取得了 78%的正确识别率。Pao 等[97]对包含生气、高兴、悲伤、厌烦和中性 5 种情感类型的 800 句汉语普通话情感语音数据库的每句语音，提取了 MFCC、LPC、LPCC、LFPC 等声学参数，采用 4 个状态的遍历离散型 HMM 取得了 88.7%的正确识别率。

6. 高斯混合模型

高斯混合模型（GMM）可以看成只有一个状态数的连续性 HMM。它使用一组加权的高斯分布来逼近特征矢量的实际分布，并根据最大似然准则进行分类决策。GMM 比较适合基于全局特征的语音情感识别。GMM 的优点是可以平滑地逼近任意形状的概率密度函数，模型比较稳定，参数比较容易处理，但 GMM 模型的阶数较难确定，缺乏理论指导，一般需要通过多次实验才能确定。

Ververidis 和 Kotropoulos[88] 使用二阶的 GMM 模型，提取基频、能量和共振峰参数，在 DES 数据库上识别 5 种情感（生气、高兴、中性、悲伤、惊奇）取得的正确识别率为 55%。Schuller 等[98]比较了 HMM 和 GMM 在语音情感识别中的性能。实验中，HMM 采用 4 个状态的从左到右结构的连续型 HMM，使用基于短时帧的基频和能量的短时统计参数识别情感，而 GMM 采用 4 个高斯分布，即 4 阶 GMM 模型，采用基于整句语音的基频和能量的全局统计参数识别情感。在包含英语、德语两种语言的 5 个人的 5250 句语音的情感语音数据库中识别 7 种情感（高兴、生气、害怕、厌恶、悲伤、中性、惊奇），HMM 获得了 77.8%的正确识别率，而 GMM 获得了 88.6%的正确识别率。

7. 多分类器组合

尽管各种模式分类器都能应用到语音情感识别中，但每种分类器都有其自身的优、缺点。为了充分利用每种分类器的优势，采用多分类器组合的方法进一步提高语音情感识别

的性能，已经成为语音情感识别方法的研究热点。多分类器组合方法可分为 3 种类型，即串联（serial）、并联（parallel）和层联（hierarchical）。串联组合的特点是将前面分类器的输出作为后面分类器的输入，最终决策结果由后面分类器决定。并联组合的特点是各个分类器都相互独立工作，然后通过决策融合规则将各个分类器的输出结果进行综合得到最终决策结果。层联组合的特点是将各个分类器组成树状（tree）层次结构，每个树节点处的分类器融合其下属层次的所有分类器的输出结果。对于层联组合，一般都采用 1～2 级进行组合，但对于如何确定最佳的层联级别还缺乏理论指导，值得进一步研究。

Hu 等[99]提出一种 GMM/SVM 的串联组合方法，即先采用 GMM 通用背景模型（universal background model，UBM）处理提取的谱特征参数产生 GMM 超向量（supervector），然后再将 GMM 超向量当作特征参数输入到 SVM 进行训练和测试。实验结果发现，这种 GMM/SVM 串联组合的性能要优于 GMM。Morrison 等[89]提出采用无加权投票（unweighted vote）和 Stacking 两种策略，将 SVM、KNN、ANN、随机森林分类器（random forest classifier，RFC）和基于样例的分类器（instance-based classifier，IBC）5 种不同分类器进行并联组合。实验结果表明，基于 Stacking 策略组合的分类器取得了最好的语音情感识别性能。Albornoz 等[90]将 GMM、HMM 和 MLP 构成一个基于谱特征和韵律特征的 2 级层联分类器模型，取得的语音情感识别性能要比单一分类器高得多。

2.1.4　语音情感识别中的难点

语音情感识别是一个多学科交叉的课题，尽管目前语音情感识别研究已经取得了较多的成果，但还面临着诸多难题。

（1）在情感语音数据库的建设方面，如何建立一个公开且理想的自然情感语音数据库供研究者使用和比较研究成果，仍然是个研究难点。

（2）在情感声学特征提取方面，如何提高声学特征参数对噪声的鲁棒性，如何将声学特征与其他非语言学信息，如语义词汇信息、上下文语境信息等有效融合用于语音情感识别，也是个研究难点。

（3）现有的语音情感识别研究主要集中于生气、高兴、悲伤、惊奇、厌恶、害怕、中性等基本情感类型的识别，忽视了对人类也很重要的非基本情感类型，如兴趣（interest）、迷惑（puzzlement）、挫折（frustration）等精神状态种类的识别研究。因此，为了提高人工情感的表达能力，如何识别更多、更复杂的情感类型也是一个研究方向。

（4）由于情感信息具有比较强的文化性和社会性，不同国家的语言、文化及民族习惯都不尽不同，因此情感表达的方式也各不相同。对某一特定的语言进行情感信息处理研究的成果，可能并不适用于其他的语言种类。因此，对于语音情感识别方面的研究完全借鉴其他语种已有的研究成果是不切实际的。应当结合实际的文化、社会等背景，研究适合于自身语言特点的语音情感识别技术也是一个方向。

（5）情感本身是一种非常复杂的心理和生理现象，各种情感的变化不但表现在言语表情上，同时也表现在面部表情、身体姿态表情及各种生理特征上。因此，针对不同应用目的，将言语表情、面部表情、身体姿态表情以及生理特征等信息有效融合在一起，进行多模态情感识别也是今后的一个重要研究方向。

2.2　人脸表情识别的技术回顾

人类语言主要包括两种，即自然语言和形体语言。用来反映视觉信息的人类面部表情，是形体语言中非常重要的一部分，它在人与人相互之间的交流中起着至关重要的作用。面部表情可以反映出不同人的思想感情和情绪状态信息，通过分析面部表情就可以获知对方的情感变化和内心态度。

面部表情是指通过人脸的眼部、颜面以及口部肌肉的变化来表现各种情感状态。面部表情不仅是人们常用的较自然的表现情感方式，也是人们鉴别情感的主要标志。科学研究表明，人脸的不同部位具有不同的表情作用。例如，眼睛对表达悲伤最重要，口部对表达高兴和厌恶最重要，前额能提供惊奇的信号，而眼睛、嘴和前额等对表达生气最重要。在日常生活中，人生气时往往双眉倒竖，怒目圆睁，颧肌抽搐，嘴角外撇，甚至咬牙切齿。对面部表情的识别，通常称为“人脸表情识别”。

1968 年，心理学家 Mehrabian[100]的研究认为，人类情感的表达，其中 7%是通过自然语言，38%是通过一些语言自带的辅助信息，如说话时的节奏、语调高低等，另外 55%则是通过面部表情表达的。1971 年，Ekman 等[101]确定了人类六种基本的人脸表情，即高兴、悲伤、惊奇、愤怒、厌恶和害怕，并系统地建立了一个含有 1000 幅图像样本的人脸表情数据库。虽然有些心理学家对这 6 种基本情感的说法有些质疑，但在计算机领域这种针对基本情感的分类已经得到大多数研究者的认可，并广泛地应用于人脸表情识别。随后，Ekman 等[102]对这 6 种基本表情的特点进一步做了定量分析，并开发出一种面部运动编码系统（facial action coding system，FACS）。在该系统中大约有 44 个既相互关联又相互独立的运动单元（action unit，AU）被用来描述人脸表情的动态变化，并详细地分析了这些运动单元 AU 的运动特征，以及它们所控制的主要区域与各种表情之间的相互关系。1991 年，Kenji[103]提出了一种采用光流法来识别人脸表情的方法。该方法是首次采用计算机技术和图像处理技术实现人脸表情自动识别的方法，受到研究者的普遍关注。

当前，人脸表情识别的研究主要侧重于人脸表情数据库的建设、人脸表情特征提取和人脸表情识别方法 3 个方面[104, 105]。20 世纪 90 年代初，国外一些著名的研究机构，如美国麻省理工学院、英国帝国理工学院等，对人脸表情识别的研究开始变得异常活跃。国内的研究始于 20 世纪 90 年代末，近年来开展的研究单位比较多，如中国科学院计算技术研究所、清华大学、浙江大学等都对人脸表情识别进行了研究。

在人脸表情识别方面，具代表性的研究工作主要包括美国麻省理工学院的 Essa 和 Pentland[106]、英国帝国理工学院的 Pantic[107]、中国科学院计算技术研究所的山世光和陈熙霖[108]、清华大学的章毓晋[109]、中国科技大学的王上飞[110]、复旦大学的姜育刚[111]、浙江大学的韦巍[112]、东南大学的郑文明和邹采荣[113]、北京航空航天大学的毛峡[114]、北京交通大学的阮秋琦[115]等。

2.2.1　人脸表情数据库

人脸表情数据库是进行人脸表情识别研究的必要条件。因此，建立一个丰富并有效的人脸表情数据库，将对人脸表情识别方面的研究具有非常重要的意义。尽管人脸表情识别方面的研究已经取得了较大的进步，但是能够获得研究者广泛认可并具有国际权威性的人

脸表情数据库仍然比较少。

目前，国内外常用的人脸表情数据库的情况如下。

1. JAFFE 数据库

JAFFE 数据库[116]是以 7 种基本表情（生气、厌恶、害怕、高兴、中性、悲伤、惊奇）为基础的数据库，包括 10 位日本女性，每种表情有 3～4 幅灰度图像，总共 213 幅图像。每幅图像的分辨率大小为 256×256 。图 2.1 给出了 JAFFE 数据库中几张表情样本图像。

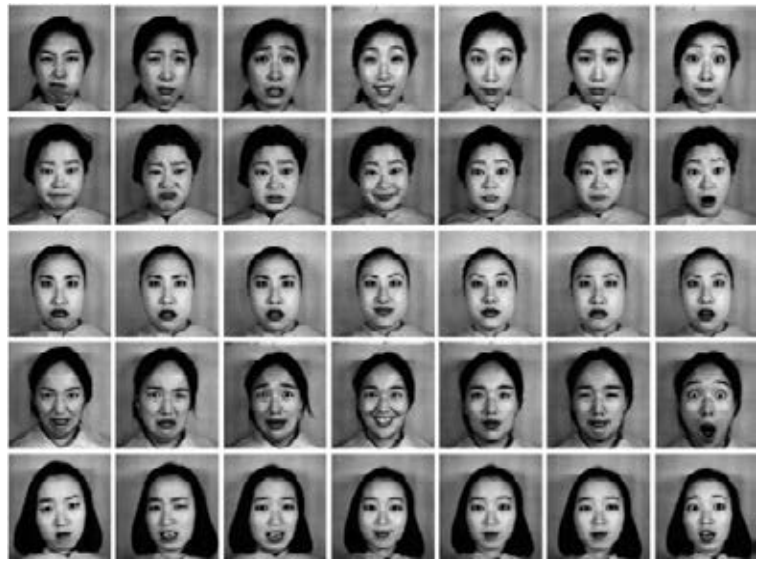

图 2.1　JAFFE 数据库中的表情样本图像

2. Cohn-Kanade 数据库

Cohn-Kanade (CK)数据库[117]是 2000 年由美国卡内基•梅隆大学中的机器人研究所和心理学系共同建立的一个基于 AU 编码的人脸表情数据库。它含 97 个对象的 486 个灰度图像序列。其中每个图像序列都包含了从中性到各种表情的过程及单个 AU 或者 AU 组合。每幅图像的分辨率大小为 640×490 。这 97 个被采集的对象年龄为 18～30 岁，其中女性占据 65%。每个被采集的对象包含 7 种基本的表情，即生气、厌恶、害怕、高兴、悲伤、惊奇和中性。Cohn-Kanade 数据库的表情样本图像如图 2.2 所示。2010 年，CK 数据库得到了进一步扩展和完善，发展出一个扩展版本 the Extended Cohn-Kanade (CK+)[118]。这个数据库包括 123 个对象和 593 个灰度图像序列。目前，CK 和 CK+数据库是人脸表情识别中比较流行的数据库。

3. MMI 数据库

MMI 数据库[119]是由英国帝国理工学院的人机交互研究实验室所创建的，它包括 19 个对象的 740 幅静态图像以及 848 个动态的图像序列。静态图像均是 24 位真彩色图像，分辨率大小为 720×576 ，而图像序列都是采用 PAL 制式，以 24 帧/s 采集所获得的。被采集对象的年龄为 19～62 岁不等，其中女性占据了被采集对象中的 44%。

图 2.2　Cohn-Kanade 数据库中的表情样本图像

4. 其他人脸表情数据库

SFEW 数据库[120]是从自然情感 AFEW 数据集中抽取的有表情的静态帧，共 700 幅图像。表情类型有 7 种，即生气、高兴、厌恶、惊奇、悲伤、害怕及中性。

FGnet 数据库[121]包括 19 个对象的 7 种表情的图像序列。其中，每个人的每种表情都有 3 个图像序列，总共 399 个图像序列。每幅图像都是 24 位彩色图像。每个图像序列的长度为几十帧到几百帧不等。

此外，还有一些常见的人脸数据库也包含了部分非基本的表情。例如，CMU PIE[122]人脸数据库则含有眨眼等表情，Yale[123]人脸数据库则含有眨眼、困乏等表情，AR[124]人脸数据库则含有尖叫等表情。

国内研究者近几年也建立了一些数据库。例如，CAS-PEAL[125]人脸数据库是由中国科学院计算技术研究所建立的，它包含了惊奇、中性、微笑、皱眉、闭眼和张嘴等表情。北京航空航天大学的毛峡等[126]建立了一个名为 BHU 的人脸表情数据库，它包括三类人脸表情，即 18 种单一表情、3 种混合表情以及 4 种复杂表情。

2.2.2　人脸表情特征分析

人脸表情特征分析主要包括两个步骤：一是表情特征提取；二是表情特征降维。

1. 表情特征提取

原始表情特征的提取方法主要有四大类，即形变特征提取法、统计特征提取法、运动特征提取法、模型特征提取法。其中，形变特征提取法是将人脸面部的一些特殊形变信息，如纹理变化或几何形变提取出来；统计特征提取法是采用统计法对人脸表情图像的特点进行描述；运动特征提取法是将某些特征区或特征点的运动信息，如特征区的光流变化或特征点的运动距离提取出来；模型特征提取法是以人脸图像中人脸的形状与纹理结构信息为基础，构建二维或三维的模型，然后通过调节模型参数的变化来匹配人脸图像中的人脸部分，最后这些模型的参数就当作是所提取的模型特征，用于人脸表情识别。

上述每一种表情特征提取方法的具体情况如下所述。

1）形变特征提取法

人脸的形变可以用几何形变和纹理变化来描述。几何形变是指人脸表情发生变化所引起的一些特征点之间的几何相对距离的改变。纹理变化则是指人脸表情发生变化所引起的一些面部纹理信息的出现或消失等。比较有代表性的形变特征提取法有特征基准点、Gabor小波变换和局部二元模式（local binary patterns，LBP）。

Pantic 等[127]提出采用检测器进行人脸检测，然后从人脸的侧面提取出 10 个特征点，正面再提取出 19 个特征点，如图 2.3 所示。利用这 29 个特征点位置的变化信息，可以识别出 32 个运动单元 AU 的运动情况，最后根据运动编码系统（FACS）识别出不同的面部表情。

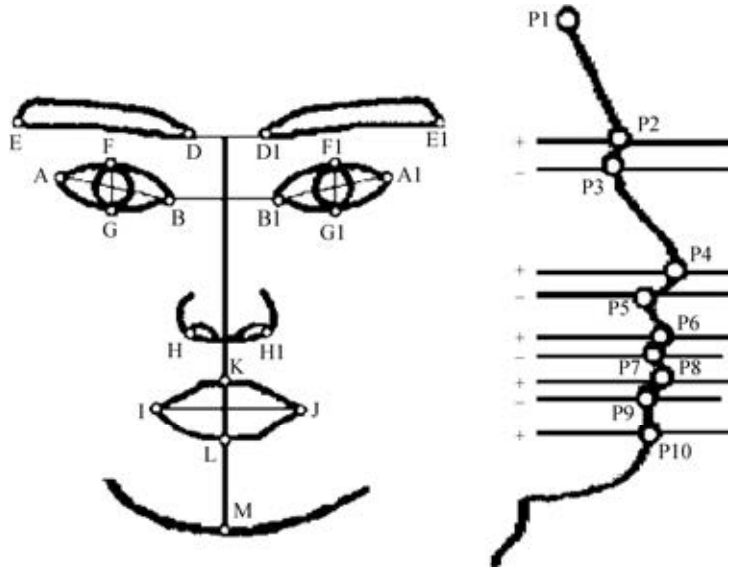

图 2.3　人脸正面和侧面选取的 29 个基准点[127]

Zheng 等[128]人工标记出人脸正面的 34 个点，如图 2.4 所示，然后将这 34 个点的几何信息转化成标号图（labeled graph）用于人脸表情识别。这些特征基准点一般都能较好地反映人脸面部肌肉的运动变化信息。基于特征基准点提取的特征向量的维数一般比较低，计算量也相对较小。然而，对于这些特征基准点的确定，一般都是通过人工辅助标记完成的，目前还缺乏精确的自动标记技术。

图 2.4　人脸正面选取的 34 个基准点[128]

Gabor 小波变换的优点主要在于它能做到对频域信号和时域信号的局部化，在频域和空域同时进行测量，因此具有良好的频率特性和方向选择特性。此外，它能容忍一定程度上的图像变形和旋转，而且对人脸姿态和光照的变化不太敏感，因此 Gabor 小波变换已经被广泛应用于人脸表情特征的提取。它的不足之处就是 Gabor 提取的特征维数较高，计算量较大，运算时间比较长，因此经常需要与其他的特征降维方法配合起来一起使用。

Zhang 等[129]提出采用多尺度和多方向的 Gabor 小波来对人脸正面选取的 34 基准点作 Gabor 小波变换，然后以 Gabor 变换之后的幅值作为表情特征用于人脸表情识别。图 2.5 给出了 Gabor 小波对两幅人脸表情图像进行特征变换的结果。

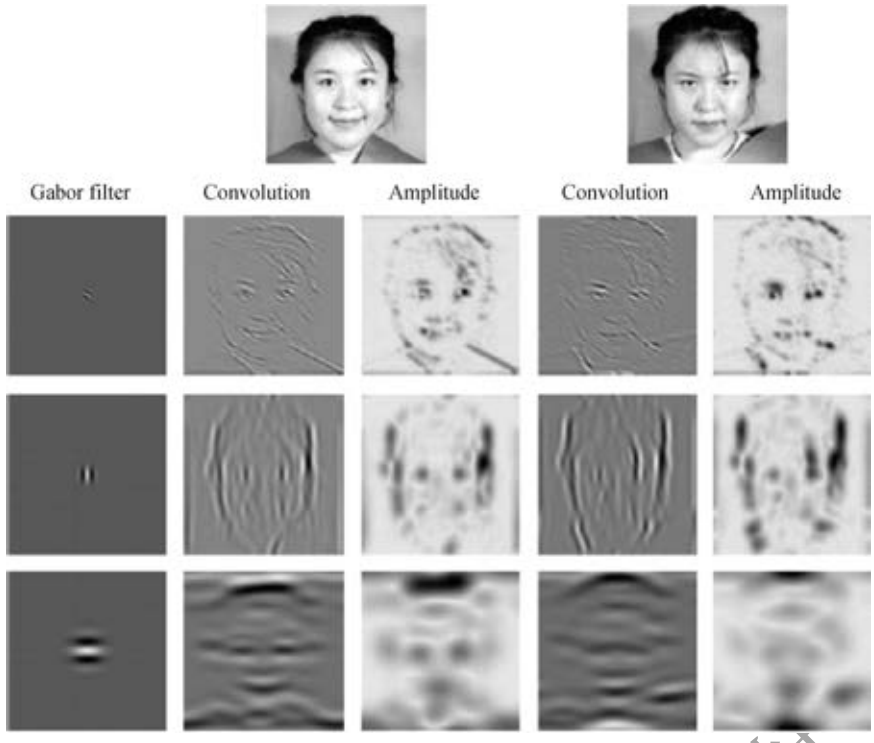

图 2.5　Gabor 小波的特征变换[129]

局部二元模式（LBP）最初是由 Ojala 等[130]于 1996 年率先提出的，并成功将 LBP 算子应用于纹理的分类[131]。LBP 本质上是一种非常有效的局部纹理描述算子，因为 LBP 可用来提取和度量灰度图像中的局部邻域之间的纹理信息。它通过对一幅图像中的每一个像素值与其邻域内的像素灰度值的大小相比较，然后用二进制数值模式来表示像素比较的结果，以达到描述图像纹理信息的目的。

LBP 算子的优点是，具有良好的旋转以及灰度的不变性，而且能够克服图像的位移、旋转以及光照不均衡等方面的问题。其计算比较简单，并且能够有效提取代表图像本质的纹理特征信息。基于上述优点，LBP 算子最初是用来做纹理分析的。近年来，LBP 算子也被广泛应用于人脸表情中的特征提取。

Shan 等[132, 133]在对人脸表情图像进行预处理之后，采用均匀的 LBP 算子扫描一幅人脸图像，从而得到人脸图像的 LBP 编码，紧接着计算出 LBP 直方图，最后将所有提取的直方图组合成一个向量作为人脸表情识别中的特征。

2）统计特征提取法

统计特征一般都具有一定的不变性，使用统计特征作为人脸表情特征，能够对图像的旋转、平移、尺寸变化等表现出一定程度上的鲁棒性。

Choudhury 和 Pentland[134]使用帧间差分的方式先对序列图像中的眉毛、眼睛等特殊部位进行预处理，然后对差分图像的直方图进行分析，获取人脸表情的特征描述。Shinohara 和 Otsuf[135]采用 35 种高阶的局部自相关方面的特征参数作为人脸表情特征使用。

统计特征提取法的优点是能够保留较多的图像原始信息，而且处理过程也不复杂。该方法的缺点在于，它对外在因素以及表情本身带来的图像差异性不加任何区分，因此该方法对人脸表情图像的预处理方面的要求较高。例如，对输入图像要进行灰度归一化，以便减小光照的影响；对人脸尺寸也要作归一化，以便消除不同人脸形状差异方面的影响等。

3）运动特征提取法

运动特征提取法主要用于对动态表情图像序列的特征提取。提取动态表情图像序列的运动特征信息，最常用的方法就是光流法（optical flow）。

光流法是一种有效处理动态图像序列的方法，其基本思想是把运动图像函数作为一个基本的函数，然后利用图像强度守恒原理来建立一个光流约束方程，最后求解出该光流约束方程得到运动参数。光流能够反映点的运动信息，而人脸表情发生变化时，主要体现在眉毛、眼睛、嘴巴等各点的弯曲、上下等变化。因此，利用光流法就可表示出人脸表情发生过程中不同的表情特征点的运动信息。

Otsuka 和 Ohya[136]采用光流法对人脸的一些特殊区域，如眼睛和嘴巴的运动进行了分析。Essa 和 Pentland[106]采用光流法对人脸面部运动进行估计，而且提出一种面部肌肉运动模型来描述人脸面部的运动。He 等[137]利用人脸的先验知识信息，对传统的光流法进行改进，并成功应用于人脸表情的识别，取得了较好的识别效果。

光流法的优点在于它能较好地反映出人脸表情的变化，而且它受光照不均匀性方面的影响比较小。其缺点在于，该方法计算量比较大，特征提取时容易受到脸部非刚性运动的影响。

4）模型特征提取法

最具代表性的模型特征提取法是活动外观模型（active appearance model，AAM）[138]方法。AAM 方法本质上是求解一个最优化问题，即利用图像及模型参数合成一个表观模型，然后通过调整模型参数，使得实际图像与模型表观的差别达到最小化。采用 AAM 方法提取人脸表情图像的特征，其过程包含两个步骤，即 AAM 的生成和采用 AAM 方法进行人脸表情特征的提取。

AAM 的生成是指把 AAM 的表观模型以仿射变换的形式将之映射到相对应的形状实例中去，从而得到一个表示当前对象的 AAM 模型。AAM 通过对目标对象的变化程度进行参数化描述。该过程是，首先从已给的训练数据集中提取对象的模型，然后利用某种规则对模型再进行组合，从而形成一个新的对象。这种方法能够在新的图像中搜索定位的既定目标对象，并且使用较少的参数对新的对象进行描述。AAM 建立的模型中包含了对象的形状和纹理两种属性的描述。在 AAM 拟合的过程中，要实现的目的是通过采用相应的拟合算法，不断地调整新生成的 AAM 模型实例与待定位的目标对象进行再匹配，直到新生成的模型实例能够与该对象达到相匹配的程度。因此，AAM 模型实例的生成是 AAM 方法中的一个重