虚拟交互环境中情绪识别研究

薛彤

2022 年 6 月

中图分类号： TP391

UDC分类号： 004

虚拟交互环境中情绪识别研究

作 者 姓 名 薛彤

学 院 名 称 计算机学院

指 导 教 师 丁刚毅教授

答辩委员会主席 吴中海教授

申 请 学 位 工学博士

学 科 专 业 软件工程

学 位 授 予 单 位 北京理工大学

论 文 答 辩 日 期 2022 年 6 月

Research on Emotion Recognition in Virtual Interactive Environments

Candidate Name:	Tong Xue
School or Department:	Computer Science and Technology
Faculty Mentor:	Prof. Gangyi Ding
Chair, Thesis Committee:	Prof. Zhonghai Wu
Degree Applied:	Doctor of Philosophy
Major:	Software Engineering
Degree by:	Beijing Institute of Technology
The Date of Defence:	June, 2022

虚拟交互环境中情绪识别研究

北京理工大学

研究成果声明

本人郑重声明：所提交的学位论文是我本人在指导教师的指导下进行的研究工作获得的研究成果。尽我所知，文中除特别标注和致谢的地方外，学位论文中不包含其他人已经发表或撰写过的研究成果，也不包含为获得北京理工大学或其它教育机构的学位或证书所使用过的材料。与我一同工作的合作者对此研究工作所做的任何贡献均已在学位论文中作了明确的说明并表示了谢意。

特此申明。

作者签名：蔚彬　　签字日期：2022年6月13日

关于学位论文使用权的说明

本人完全了解北京理工大学有关保管、使用学位论文的规定，其中包括：① 学校有权保管、并向有关部门送交学位论文的原件与复印件；② 学校可以采用影印、缩印或其它复制手段复制并保存学位论文；③ 学校可允许学位论文被查阅或借阅；④ 学校可以学术交流为目的,复制赠送和交换学位论文；⑤ 学校可以公布学位论文的全部或部分内容（保密学位论文在解密后遵守此规定）。

作者签名：蔚彬　　导师签名：

签字日期：2022年6月13日　　签字日期：2022年6月13日

摘要

虚拟现实带来复杂真实情境的交互式视景仿真，在军事、教育、医疗等领域具有广阔的应用前景。情绪在人的认知、决策、社交等过程中起着重要作用，识别并理解人在虚拟交互环境中的情绪状态，能够从根本上改进人与技术交互方式，提供更好的用户体验。情绪识别旨在定义、测量人的情绪状态，建立生理信号、行为动作等模态信息与情绪之间的映射关系。由于人类情绪的主观性与复杂性，以及人在虚拟环境中交互行为的自由性与多样性，研究的关键科学问题是：如何获取虚拟交互环境中精确有效的情绪 Ground-Truth 标签，突破面向视觉交互行为的情绪识别难题。本文结合虚拟现实、人机交互与情感计算相关研究，从理论层面构建了虚拟交互环境中情绪识别概念模型及情绪识别系统，聚焦多模态、细粒度情绪识别研究方法。主要完成的工作与创新点如下：

（1）实时连续情绪测量。针对虚拟交互环境中现有情绪标注方法不实时不连续、耗时长且干扰用户体验等问题，提出了基于“唤醒-效价”二维情绪模型的 HaloLight 与 DotSize 两种情绪标注信息可视化方案，并给出了相应的情绪诱发及测量实验范式。研究构建了一个实时连续情绪标注方法可用性评估框架，从用户体验质量和标注数据有效性两个方面进行验证。实验表明本文提出的方法能够在不干扰用户虚拟体验的前提下，获取精确有效的细粒度情绪 Ground-Truth 标签。

（2）多模态情绪数据集构建。针对虚拟交互环境中多模态情绪数据集空白问题，本文提出一个公开的生理及行为多模态连续情绪数据集 CEAP-360VR，包含用户在虚拟体验中的视觉行为数据、生理信号、实时连续情绪标注及主观报告量表等内容。采用统计学及机器学习方法从多个角度验证了数据集的有效性与可信度。该数据集为虚拟交互环境中的情绪识别研究提供了良好的数据源；实现了在细粒度层级上开发、验证情绪识别算法，构建用户在虚拟体验中更精确的情绪识别模型。

（3）视觉交互行为与情绪的相关性研究。针对虚交互环境中用户视觉行为的复杂性，本文重点关注头部运动与眼部运动两种行为要素，从用户之间与用户自身的视觉行为关系及统计学偏向两个角度分析行为特征。研究首次提出了一种片段层级的头部运动、眼部运动及特征与连续情绪标签之间的细粒度相关性识别方法，实验表明二者之间具有显著相关性。研究实现了一种低成本方法，基于隐式行为信息动态调整呈现的情绪内容，提高沉浸式虚拟交互环境中的用户体验质量。

（4）基于视觉交互行为的情绪识别研究。引入了一个新问题——如何面向复杂视觉行为构建用户独立的情绪 Ground-Truth 标签。针对虚拟环境中个体之间的反应延迟及交互行为的多样性，提出了基于情绪诱发内容参考特征的实时连续情绪标注序列时间对齐方法、及基于视口的连续情绪融合方法。实验表明该方法融合后的连续情绪序列准确有效，且提供了情绪状态的峰值、波谷、变化趋势等时序性细节信息，实现了在细粒度层级上理解用户情绪状态与对应诱发情境之间的关联性。

关键词：虚拟交互环境；连续情绪 Ground-Truth；多模态情绪数据集；视觉交互行为；“唤醒-效价”二维情绪模型

Abstract

Virtual reality brings interactive visual simulation of complex real situations, and has broad application prospects in military, education, medical and other fields. Emotion plays an important role in people's cognition, decision-making, and social interaction. Identifying and understanding people's emotion states in virtual interactive environments can fundamentally improve the way people interact with technology and provide better user experience. Emotion recognition aims to define and measure people's emotion states, and to establish the mapping between modal information such as physiological signals, behavior, and emotion. Due to the subjectivity and complexity of human emotion, as well as the freedom and diversity of human interaction in virtual environments, this research aims on: how to obtain accurate and effective emotion Ground-Truth labels in virtual interactive environments, conducting emotion recognition based on interaction. This paper combines the related research of virtual reality, human-computer interaction and affective computing, constructs the conceptual model and system of emotion recognition in virtual interactive environment from the theoretical level, and focuses on multi-modal and fine-grained emotion recognition research methods. The main work and innovations are as follows:

(1) Real-time continuous emotion measurement. Aiming at the problems that the existing emotion annotation methods in virtual interactive environments are not real-time and dis-continuous, time-consuming and interfere with the user experience, we proposed Halo-Light and DotSize, two emotion annotation information visualization schemes based on the "Arousal-Valence" two-dimensional emotion model. The corresponding experimental paradigm of emotion induction and measurement is given. The research constructs a real-time continuous emotion annotation method usability evaluation framework, and verifies it from two aspects of user experience quality and validity of annotation data. Experiments show that the method proposed in this paper can obtain accurate and effective fine-grained emotional Ground-Truth labels without disturbing the user's virtual experience.

(2) Construction of a multi-modal emotion dataset. Aiming at the blank problem of multi-modal emotion dataset in virtual interactive environments, we propose a public physiological and behavioral multi-modal continuous emotion dataset CEAP-360VR, which in-

cludes visual behavior data, physiological signals, and real-time continuous emotion annotation during virtual experience, as well as post-stimuli subjective questionnaires. Statistical and machine learning methods are used to verify the validity and reliability of CEAP-360VR dataset from multiple perspectives. This dataset provides good data source for emotion recognition research in virtual interactive environments. It can develop and verify fine-grained emotion recognition algorithms, and build more accurate emotion recognition models for virtual environments.

(3) Research on the correlation between visual behavior and emotion. In view of the complexity of user visual behavior in virtual interactive environments, this paper focuses on head movement and eye movement, and analyzes the behavior characteristics from the visual behavior relationship among users and the user themselves and statistical bias. Our study first proposes a fine-grained correlation identification method between segment-level head movement, eye movement, and features and continuous emotion labels. Experiments show that there is a significant correlation between them. Our research implements a low-cost method to dynamically adjust the presented emotional content based on implicit behavior information to improve the quality of user experience in immersive virtual interactive environments.

(4) Research on emotion recognition based on visual interaction behavior. We introduce a new problem —how to construct subject-independent emotion Ground-Truth labels for complex visual behavior. Aiming at the response delay and diversity of interaction among individuals in virtual environments, we propose a time-alignment method based on reference features for real-time continuous emotion annotation sequences, and a viewport-dependent continuous emotion fusion method. Experiments show that the continuous emotion fused by this method is accurate and effective, and provides time-series detail information such as peaks, valleys, and changing trends of emotion states. Our research contributes to better understanding of the relationship between user emotion states and corresponding evoked situations at a fine-grained level.

Key Words: Virtual Interactive Environments; Continuous Emotion Ground-Truth; Multimodal Emotion Dataset; Visual-based Interaction；”Arousal-Valence” Emotion Model

目 录

插 图

表 格

第 1 章　绪论

1.1　研究背景

随着智能媒体技术（Intelligent Media，IM）和头戴显示器设备（Head-Mount Display，HMD）的迅速发展，虚拟现实（Virtual Reality，VR）带来复杂真实情境的交互式视景仿真[1]。人类通过视觉[2]、听觉[3]、触觉[4] 和嗅觉[5] 等多通道感知虚拟场景信息；同时借助语言[6]、手势[7]、身体动作[8] 和外部辅助设备[9] 等作为输入信息，与虚拟环境及其中的元素进行交互。如何理解人在虚拟交互体验中的行为与认知，是 VR 研究的一个核心问题[10]。在人的决策、感知、记忆、社交等过程中，情绪起着至关重要的作用，同时还影响着人的生理心理状态[11]，这使得情绪识别（Emotion Recognition，ER）成为虚拟交互环境中最基本的研究课题之一。

1985 年，图灵奖获得者 Minsky 教授率先提出借助计算机识别人类情绪的想法，在其专著《The Society of Mind》[12] 中指出“问题不在于智能机器能否拥有情感，而在于没有情感的机器能否实现智能”，强调了计算机学习分析情绪的必要性和重要性。1997 年，美国麻省理工学院媒体实验室的 Rosalind Picard[11] 首次提出情感计算（Affective Computing，AC）这一概念，即“通过赋予计算机识别、理解、表达和适应人的情感的能力来建立和谐人机环境，并使计算机具有更高、更全面的智能”。情绪识别是情感计算的重要环节，识别人在虚拟体验中的情绪状态并作出适当反馈，有可能从根本上改变人类与技术交互的方式[13]，让虚拟环境的交互更加自然。情绪作为人内部的一种复杂体验[14]，具有如下特点：（1）实时产生且连续变化[15,16]：人的情绪经由所处情境或事件诱发实时产生，在时间维度上是动态连续变化的；（2）伴有生理心理及行为变化[17,18]：人在不同的情绪之下会表现出不同的表情动作，人体器官与内部组织也会产生一系列的生理心理变化；（3）个体特殊性[19,20]：由于个体之间的文化背景、教育经历、观念喜好等各不相同，同一诱发内容所引起的情绪状态也会存在差异。此外，虚拟环境还带给人更自由地交互行为模式[10]。因此虚拟交互环境中精准有效的情绪识别仍是一项具有挑战性的研究工作。

为了识别和理解情绪，首先应该对其进行量化建模。现有的情绪表示方法主要基于两类情绪模型[21]：离散的情绪模型和基于维度的情绪模型。离散情绪模型采用离散标签（开心、恐惧、悲伤等）描述情绪，是一种简单直观的情绪表示方法[22,23]；但是，

该方法无法衡量情绪强度，也没有考虑情绪体验随时间变化的动态连续特性[21]。基于维度的情绪模型采用 N 维（$N \geq 2$）空间描述情绪状态，最常见的是 Russel 等人[24]提出的“唤醒-效价”二维情绪模型（Circumplex Arousal-Valence Emotion Model）。相比于离散情绪模型，基于维度的情绪模型能够描述更广泛的细粒度情绪（Fine-grained Emotion）[25]；不同于在一段体验中识别一种情绪，细粒度层级的情绪识别捕获了情绪的时变特征，对情绪的时间动态性进行建模，在时间尺度上更加精确[21,26]。

相比于传统的电脑端和移动端媒介，虚拟环境提供了更具交互性的沉浸式体验，其可控的仿真实验情境逐渐成为心理学等基础科学研究的一种方法论工具[27–29]。无论是在虚拟环境中开展军事训练[30]、模拟课堂教学[31]、或是观看全景视频[32,33]，识别整个沉浸式体验中人的情绪状态至关重要[28]。在学术研究方面，人机交互、智能媒体领域的国际重要会议、期刊如 ACM CHI、ACM MM、IEEE TMM、IEEE TAFFC 等每年都有关于人的情绪识别研究成果发表；虚拟现实国际顶级会议 IEEE VR 开设了专门的用户多感官体验主题，以及“情绪与感知（Emotion and Cognition）”分会议题[34]。以人为中心的情绪识别通常基于人在虚拟体验中的生理信号[35–37]、行为动作[2,38] 等可测量指标。然而，认知科学家 Lisa Feldman Barrett[39] 指出，对于用户主观、内部的情绪状态，只能通过自我报告方式获取，没有客观、外部的测量标准。Ei Ali 等人[13,40] 连续两年在人机交互顶级会议 CHI 上举办“瞬时情绪的诱发与获取（Momentary Emotion Elicitation and Capture，MEEC）”研讨会，探讨如何在不影响用户体验的情况下，诱发并实时捕捉有效的情绪标准数据（Ground-Truth）。

可见，在虚拟环境中开展情绪识别研究，需要采用基于维度的模型对情绪进行量化建模，构建科学可信、且符合道德准则的虚拟交互环境诱发情绪状态；并借助人机交互和情感计算相关理论，开展精确有效的情绪测量与情绪理解研究。当前，该领域主要存在以下挑战：

（1）情绪 Ground-Truth 标签的有效获取：情绪 Ground-Truth 标签是指人类自身真实感受到的情绪状态，是情绪识别中分类模型和推理算法的训练数据。标签的精确程度直接影响情绪识别系统的可靠性和识别结果的准确度。由于情绪的主观性和复杂性，如何获取情绪 Ground-Truth 标签并评估其有效性，是虚拟环境中情绪识别的难点之一。

（2）多模态情绪数据集的构建验证：情绪识别研究离不开人类行为、生理、心理等情绪相关的多模态用户数据，这些数据的测量与理解是一个漫长、困难且昂贵的过

程。高质量公开数据集能够避免这一过程，同时还可以提升研究的可比较性与可重复性。由于不同时间分辨率的多模态数据收集、同步与验证仍是一项挑战，构建可靠的多模态数据集是虚拟环境中情绪识别亟需解决的问题。

（3）面向交互行为的情绪识别：人的认知与行为之间具有某种关联，是心理学领域的共识。交互性是虚拟现实的重要特征，人的交互行为会受到情绪影响，同时能够反应若干情绪状态。虚拟环境中人的行为具有自由性和多样性，目前尚没有统一的交互标准。因此，如何定义虚拟环境中的交互行为特征、结合交互行为进行情绪识别是当前研究的一个难点。

1.2 研究意义

虚拟交互环境中人的情绪识别研究具有十分重要的学术意义和应用价值，如图1.1。一方面，精确有效的情绪识别有助于推进对虚拟环境中人的行为及认知探索研究；另一方面，虚拟现实的应用领域涉及军事仿真、教育教学、医学诊疗、市场营销、娱乐游戏等行业，目前已经形成了一个强大的跨学科社区[41]。

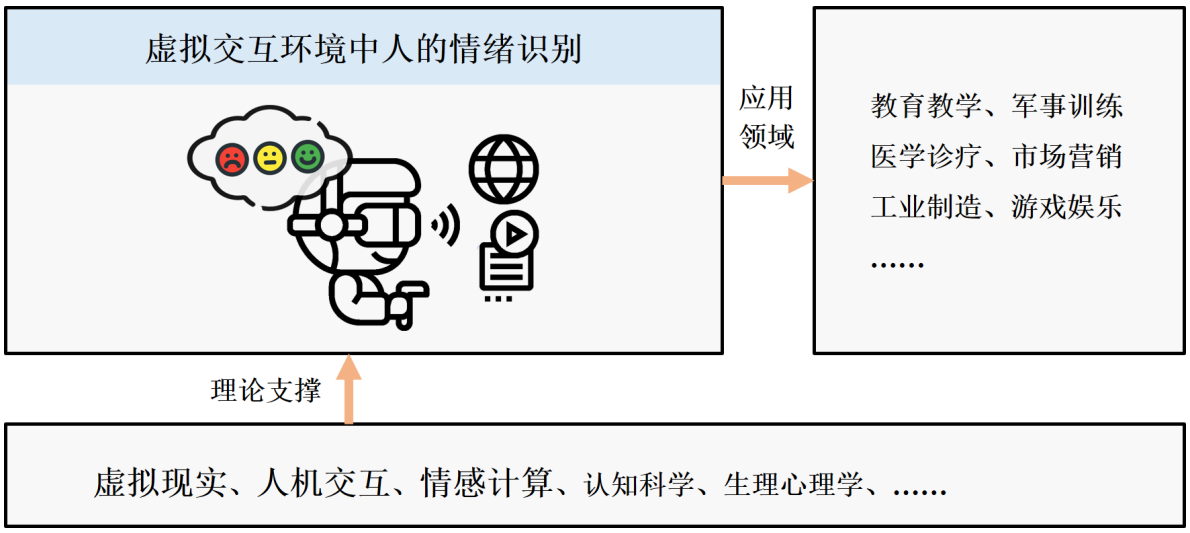


图 1.1 本研究相关的主要学科领域和应用领域

1.2.1 理论意义

虚拟交互环境中的情绪识别涉及到人、环境和可穿戴设备等多项内容，是虚拟现实与人机交互、情感计算等多个学科的交叉，还与认知科学、生理心理学等密切相关。由于虚拟环境特有的行为交互性，研究更需要强调细粒度层级的情绪识别及其与交互

行为特征之间的关系[13]。

从人机交互的角度来讲，精确理解虚拟交互环境中用户的情绪状态，一方面能够帮助研究者们设计更好的用户体验，提升虚拟现实中交互技术的可用性、可接受性和可访问性；另一方面，能够让机器更好地适应人的操控能力和情绪状态，有助于人与虚拟环境之间更加自然高效的信息交换[10]。从情感计算的角度来讲，公开可用的多模态情绪数据集能够丰富虚拟环境中情绪研究数据源，对于测试情绪识别研究问题、假设、算法或设计新的机器学习模型非常重要[17]；此外，细粒度层级的情绪识别关注瞬时情绪状态，连续情绪 Ground-Truth 标签能够启发虚拟环境中细粒度情绪识别模型的构建、优化和推理[42]。从虚拟现实的角度来讲，通过测量用户的多模态生理信号、行为数据及实时连续的情绪标注数据，有助于研究人员更全面理解人在虚拟环境中的认知状态；本文还在细粒度层级上开展面向行为特征的一系列情绪研究，为探索交互机制和虚拟空间本质提供了启示和验证手段，也为虚拟现实领域的情绪研究提出了新的机遇与挑战。

1.2.2 应用价值

情绪对人的学习、记忆、推理、决策等认知过程具有重要影响，虚拟环境通过还原各种真实情境，允许在完全实验可控的环境中捕捉人的情绪、生理及行为反应。伴随着 VR 技术和智能媒体的迅速发展，虚拟交互环境中人的情绪研究具有很大的应用前景，本节着重从教育训练、医学诊疗和市场营销三个方面介绍：

在教育训练领域，Kavanagh 等人[31] 指出 VR 具有强大的教育优势，为学生和教师提供了一个丰富的实践空间和与之互动的机会；一些研究[30] 指出 VR 技术可以构建情绪化的学习情境，诱发并监测学生情绪状态以及师生之间的情绪相似度；应用场景还可扩展至驾驶和飞行仿真模拟训练中[43,44]，VR 技术低成本、安全的训练环境能够显著提高用户的学习动机和学习成果，对实战成绩带来积极影响。在医学诊疗领域，VR 近年来已被纳入临场和非临床的情绪调解干预治疗中，通过模拟患者害怕面对的情境，与不同化身互动并执行一系列情绪识别任务，帮助其接受并改善自己的焦虑、害怕等负面情绪；另一方面，VR 在治疗患有心理或行为障碍患者（如恐惧症、焦虑症等）方面能够作为社交技能培训平台，通过测量和理解患者情绪状态促进更好的沟通[45?]。在市场营销领域，情绪体验是消费前阶段消费者行为意向的影响因子和预测因子，用户情绪反应用于评估促销广告有效性有着重要意义[46,47]；VR 技术能够在实

验环境中分析消费者的情绪体验，帮助企业有针对性的设计和测试产品属性，评判产品未来能否成功应用于市场[48]。

1.3 研究现状

1993 年，Grigore Burdea 与 Philippe Coiffet 在著作《虚拟现实技术》[49] 中指出，虚拟现实具有三个最突出的特征：（1）交互性（Interactivity）：指的是用户对虚拟环境及其中要素的可操作程度与获得反馈的自然程度；（2）沉浸感（Immersion）：指的是用户通过交互设备与自身感知系统，置身于虚拟环境的真实程度；（3）构想性（Imagination）：指的是借助虚拟现实技术构思设计虚拟空间，使抽象概念具象化的程度。因此，研究者们借助 VR 技术能够模拟复杂真实情境，实现在可控环境中研究人的情绪及其与生理信号、行为动作等相关指标之间的关系。本节对虚拟环境中人的交互行为及情绪识别研究进行简要回顾，并指出了当前虚拟交互环境中情绪识别研究存在的问题。

1.3.1 虚拟环境中的交互行为

相比于沉浸感和构想性，交互性能够带给用户最直接的体验，是 VR 技术走向应用的核心属性[10]。虚拟现实作为新一代人机交互界面，交互也从基于鼠标键盘的 2D 图形用户界面扩展到自然用户界面中[50]。人机交互技术为虚拟交互环境提供了基于不同目的与功能的多种交互模式，本节将主要交互模式归纳为以下三种：

（1）视觉交互行为。眼睛是人感知外界信息的关键通道，视觉行为揭示了人类处理视觉信息的交互机制。头戴显示器设备是 VR 普遍采用的一种立体显示器，完全覆盖人的视觉感官通道，同时能够实时跟踪记录用户在体验过程中的头部运动[51,52]。虚拟交互空间具有 6 自由度（Degree of Freedom，DoF），包含位移自由度、旋转自由度两类。头部运动常用作虚拟环境中的导航工具[53]，1995 年，Slater 等人[54] 首次通过监测用户原地行走时的头部振动识别行走动作，将用户视口转向当前方向实现导航；Terziman 等人[55] 丰富了头部运动交互方式，其中水平运动表示向前移动、向上与向下的垂直运动分别表示跳跃和爬行，速度通过头部振动的振幅进行控制。眼部运动能够带来与虚拟环境更自然、直观的交互。眼部追踪也是当前虚拟现实的一项关键技术，近年来一些研究[56,57] 通过内嵌有眼动仪的 HMD 设备采集用户实时眼部运动数据。头部运动与眼部运动的相互作用较为复杂，且神经耦合，例如前庭-眼反射

（vestibulo-ocular reflex）机制[58]；Piumsomboon 等人[59] 基于此提出了一种在虚拟环境中平滑追踪眼球运动的自然交互技术 RadialPursuit；Andrist 等人[60] 指出凝视（Gaze）在日常协作交流中的重要性，并构建了一个能够理解用户凝视状态并生成凝视效果的双向凝视模型。

（2）肢体动作交互行为。肢体动作是指人的手、手臂或腿部等肢体部位有意识或无意识的运动[61]。在虚拟交互环境中，通常采用可穿戴传感器设备、计算机视觉两种方式追踪定位人在物理空间中的肢体动作信息，解析并理解其意图。其中，计算机视觉方式下，用户交互范围比较大，在交互表达上最大程度地符合人对真实环境已有的认知，交互状态更加自然。可穿戴传感器设备包括数据手套、运动捕捉设备等，能够实时捕获用户的高质量运动数据，对细节动作的识别精度更高；同时，设备自带的反馈装置可以提供力反馈等，使得用户在交互中感受到目标体的物理特性，增强真实感。Chagué 等人[8] 提出一个结合动作捕捉设备的沉浸式虚拟平台，用户能够自由地在虚拟空间中漫游，同时与空间中的三维物体或是其他人物角色交互。Zhang 等人[62] 借助 Leap Motion 设备开发了一种用户在虚拟漫游中的双手交互方法，通过前后旋转左手掌控制第一人称角色前后移动、通过右手拇指的指向控制角色左右移动，实验表明采用双手操作符合人的日常交互行为，有助于提升用户在虚拟环境中的沉浸感。

（3）外部辅助设备交互行为。人类主要凭借双手触摸物体。虚拟环境中另一种重要交互方式是采用类似于键盘鼠标的外部辅助设备。目前三大主要的 VR 头戴显示器设备提供商 SONY[27]、OCULUS[63]、HTC VIVE[9,64]，均配有 VR 手柄作为用户与虚拟空间交互的主要设备；这些手柄设备具有 6 自由度的实时追踪功能，用户通过按钮、摇杆等元件进行输入操作，还可以感受到震动等触觉反馈效果[4]。Gusai 等人[64] 与 De Paolis 等人[9] 证明了 HTC VIVE 控制器在虚拟环境任务执行中具有良好的时间性能与准确度。Lee 等人[65] 开发了一款针对虚拟环境的手持控制器 TORC，能够在手掌上产生顺应性、纹理、抓取释放虚拟体等灵巧操作的触觉，该设备有助于用户更精确地旋转和定位物体。

除了上述三种交互行为，还有语音交互[6]、多通道交互[66] 等方式。由于视觉交互是人与虚拟环境及要素交互的基础，研究[67,68] 也指出人类获取的 80% 以上信息来自视觉系统，在本文后续章节的研究中，主要考虑基于头部运动与眼部运动的视觉交互行为。

1.3.2 虚拟环境中的情绪识别

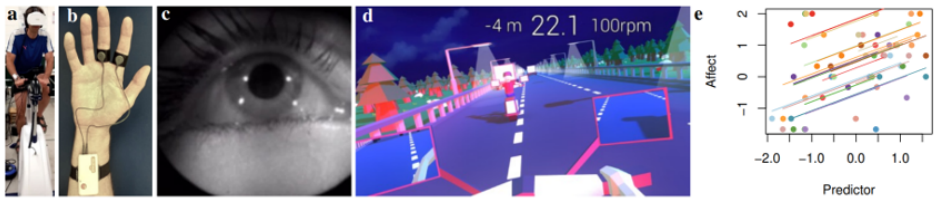
自 2000 年以来，虚拟现实越来越多地应用于情绪识别研究并受到了广泛关注。Chris Milk[69] 将 VR 定义为一个终极共情机器；相关研究[70] 利用 VR 技术构造虚拟交互环境，在可控实验环境下诱发、测量并理解用户情绪反应。虚拟交互环境给用户带来一个完全仿真的世界，其中的任何细节都有助于诱发用户情绪[71]。Riva 等人[72] 采用三种不同的虚拟公园场景分别诱发用户的中性、焦虑与放松情绪状态；Felnhofer 等人[27] 通过控制虚拟环境中的光照、天气、时间、他人面部表情等变量诱发参与者高兴、悲伤、无聊、生气与焦虑五种情绪；Naz 等人[73] 定义了虚拟环境中基于颜色和亮度获取情绪响应的设计原则。Chirico 等人[74] 借鉴真实环境中的经验，采用瀑布、山峰和太空视角能够诱发用户的敬畏感；Hedblom 等人[75] 指出相比于城市场景，虚拟环境中的自然场景能给用户带来更放松体验。Tabbaa 等人[57] 指出 VR 技术用于情绪诱发，主要有以下三个方面优势：（1）VR 通过对真实环境及其中关键要素进行仿真模拟，给用户带来身临其境的幻象；（2）相比于非沉浸环境和半沉浸环境，虚拟环境的临场感和交互性能够让用户更深入地参与到诱发素材中；（3）伴随着 HMD 设备的多样化和迅速商业化，虚拟环境更容易搭建并部署服务于终端用户。

在情感计算领域，研究者们[17] 采用各种类型的生理信号及特征，试图通过机器学习技术构建生理数据与情绪变化之间的固定关系。当前研究已经揭示了与中枢神经系统（Central Nervous System，CNS）相关的脑电信号（Electroencephalography，EEG）[76] 及与外周神经系统（Peripheral Nervous System，PNS）相关的皮肤电活动（Electrodermal Activity，EDA）[77]、皮肤表面温度（Skin Temperature，SKT）[78]、心率变异性（Heart Rate Variability，HRV）[79] 等生理信号均能够反应特定情绪状态，这些生理测量指标也逐渐应用于虚拟环境。2002 年，Jang 等人[37] 分析了 11 名用户在虚拟环境中针对驾驶和飞行模拟两项任务的生理响应，监测心率、皮肤阻力和皮肤温度三项指标，研究发现皮肤阻力和心率变异性能够反应用户的唤醒度，可用于评估情绪状态。Nasoz 等人[80] 在虚拟环境中诱发用户模拟驾驶时的负面情绪，如恐慌/害怕、愤怒/沮丧、厌倦/疲劳，并收集皮肤电、心率等生理信号，采用多媒体技术识别并自适应驾驶员情绪，提高驾驶安全性。Marín-Morales 等人[35] 提出一个基于 VR 的情绪诱发与识别系统，如图1.2(a)所示，通过脑电信号和心率变异性研究用户的唤醒、效价两个维度的情绪感知数据；支持向量机（Support Vector Machine，SVM）分类器得到的唤醒和效价识别准确率分别为 75% 和 71.21%，表明采用生理信号数据能够有效地识别用户情绪。

Bălan 等人[81] 融合 EEG、HRV、EDA 生理指标，采用一组机器学习方法 SVM、线性判别分析（Linear Discriminant Analysis，LDA）、随机森林（Random Forest，RF）等开展 VR 游戏中关于恐高症的自动情绪识别研究，识别准确率在 42.5% 至 89.5% 之间。然而，用户在虚拟体验的交互过程中通常存在频繁的头部与身体运动，这些都会影响传感器、脑电设备等生理测量工具的稳定性和可靠性[82]。

交互性是虚拟环境的重要特性。Kim 等人[83] 采用平板传感器采集用户在观看悲伤、快乐、平静三种类型 VR 视频时的步态特征，结果表明用户在开心状态下的步幅与步速会显著提升，足底压力分布峰值在第一、二跖骨区较重。Reichenberger 等人[84] 在一项针对社交恐惧的心理学实验中测量用户的眼部凝视行为，用于探索注意力对情绪状态的影响。最近，虚拟环境中人的情绪研究开始关注头部运动、眼部运动等交互行为特征与情绪的关系。Li 等人[63] 采集了 93 名被试观看 73 个全景视频时的头部运动数据和 SAM 情绪标注数据并探索二者之间关系，发现用户头部的俯仰运动与情绪唤醒维度之间具有正相关，头部运动偏航角的标准差与效价维度之间具有正相关。Barathi 等人[2] 在一项高强度游戏的情绪识别研究中收集了用户的眼部运动数据，如图1.2(b)所示，实验表明眼部的注视、眨眼、瞳孔直径与情绪的唤醒和效价维度具有相关性。最近，Tang 等人[38] 采集了 19 名被试观看静态全景图片时的唤醒和效价情绪报告值，分析表明消极情绪对眼部注视与扫视特征有明显影响，而积极和中立的情绪状态则没有显著影响。采用虚拟交互环境中用户的视觉交互行为进行情绪识别，是 VR 用于情绪研究的一个潜在强大优势。

(a) 基于生理信号的情绪诱发与识别系统[35]



(b) 基于皮肤电及眼部运动的高强度 VR 游戏中的情绪识别[2]

图 1.2　虚拟环境中基于生理及行为模态的情绪识别系统

综上，虚拟交互环境能够诱发用户广泛的情绪状态，现有的情绪识别方法主要基于生理心理学方法、以及步态等行为特征；一些研究开始探索视觉交互行为用于情绪识别，均采用传统的统计学方法，且没有系统地研究二者之间的关系。然而，情绪作为人内部的一种复杂瞬时体验，并不总会有明显的外部表现[39]；生理信号等测量方式也具有很大的局限性[16,82]。因此，虚拟交互环境中的情绪识别还需要结合用户情绪的自我报告。

1.3.3 存在的问题

虚拟交互环境中的情绪识别研究仍处于起步阶段，本文在对相关研究进行分析总结后，发现其主要存在如下三个方面的问题：

（1）实时连续情绪标注实现困难：情感计算领域将人的情绪视为连续变量，对其进行表示、标注和建模研究。虚拟交互环境能够诱发广泛的情绪，然而目前的研究工作尚无针对佩戴 HMD 时的连续情绪报告。在虚拟体验中，用户整个视口被 HMD 覆盖，如何确保用户在看不到标注设备的情况下完成情绪报告；如何确保实时连续标注数据精确可靠，真正反应到了用户体验。目前并没有研究针对这一问题提出有效的解决方案。

（2）多模态情绪数据缺乏：为了在更科学的层次上更细致地分析用户情绪特征，研究需要获取大量可再现的情绪数据支持，即经验证的情绪数据集。尽管 VR 作为一种沉浸式媒介技术已广泛用于情绪诱发，但可用于虚拟交互环境中情绪研究的多模态数据集非常稀缺。现有的研究绝大多数基于自己的实验与数据，难以复现和供其他研究使用。目前尚无针对虚拟体验的包含实时连续情绪报告、行为数据、生理信号、主观问卷的多模态情绪数据集。

（3）行为特征与情绪连结薄弱：用户佩戴 HMD 进行虚拟体验时的头部运动与眼部运动提供了大量观看信息，但这些行为数据用于情绪识别与评估的研究仍处于初级阶段。此外，虚拟环境中用户能够自由地转动头部选择视口位置，个体之间的体验内容存在差异，而用户情绪响应是由体验内容诱发的。那么，如何将多样化体验中的情绪数据进行融合，如何借助实时连续的情绪报告数据，在细粒度层级上分析行为特征和情绪的相关性，需要进一步研究。

1.4 研究内容

本文深入调研了虚拟环境中用户交互行为及情绪识别的研究现状。首先从理论层面对虚拟环境中的情绪识别概念模型、情绪识别系统及多模态、细粒度情绪识别方法进行研究，聚焦视觉交互行为与细粒度情绪识别。研究提出构建虚拟体验中精确有效的连续情绪 Ground-Truth 标签与多模态情绪数据集、突破面向视觉交互行为进行情绪识别的关键性和迫切性。主要工作包括以下四个方面：

（1）实时连续情绪标注方法及评估体系

针对虚拟交互环境中现有情绪标注方法不实时不连续、耗时长且干扰用户体验等问题，本文结合虚拟现实、人机交互和认知心理学科，聚焦虚拟环境中实时连续情绪标注方法的三个设计原则，采用高分辨率的 HMD 设备为用户提供高质量内容和带有摇杆的无线数字游戏控制器作为情绪标注设备；通过多领域专家共同设计，从情绪类型与强度等角度对标注方法进行多轮迭代评估，提出 HaloLight 和 DotSize 两种标注信息可视化方案。为了验证虚拟环境中实时连续情绪标注方法可用于构建有效的情绪 Ground-Truth 标签，研究提出一个连续情绪标注方法可用性评估体系，从虚拟交互环境中的用户体验质量和标注数据的有效性两个方面给出了评估指标及对应的评估方法。研究构建虚拟交互环境中实时连续情绪诱发及测量实验范式，评估基于 HaloLight 与 DotSize 的情绪诱发实验场景和情绪数据采集系统的可用性。

（2）生理、行为及连续情绪标注多模态数据集

多模态情绪数据集是情绪识别的关键基石。针对虚拟交互环境中多模态情绪数据集空白问题，本文构建一个公开的实时连续生理和行为情绪标注多模态数据集（Continuous Physiological and Behavioral Emotion Annotation Dataset for 360° Videos，CEAP-360VR）[1]，包含 32 位被试观看八个全景视频的诱发素材与体验后问卷数据，观看过程中实时采集的连续情绪标注数据、头部与眼部运动信息、瞳孔直径数据、外周生理信号，以及用户数据获取、处理和验证脚本。为了帮助研究人员更好地复现及使用 CEAP-360VR 数据集，研究对多模态用户数据进行预处理和统计学分析，从多个角度分别验证实时连续标注数据、瞳孔直径数据、外周生理信号的有效性；借助机器学习技术进行一系列分类基线实验，进一步验证 CEAP-360VR 数据集的有效性和可信度。研究为虚拟交互环境中的情绪识别提供了良好的数据源，极大地丰富了细粒度层级上情绪测量与理解研究。

[1]数据集链接：https://github.com/cwi-dis/CEAP-360VR-Dataset

（3）视觉交互行为与情绪的相关性研究

理解用户在虚拟环境中的视觉特征与探索模式非常重要。本文主要关注用户佩戴HMD时的头部运动与眼部运动两个关键视觉行为要素，以及从中提取的注视、眼跳等行为特征值；研究从用户之间与个体自身的视觉交互行为关系、及视觉行为的统计学偏向两个角度出发，结合人的视觉注意机制相关研究分析虚拟环境中常见的四种视觉行为特征。为了进一步理解虚拟环境中视觉交互行为与情绪间的关系，研究提出一种片段层级的用户头部运动、眼部运动及特征与连续情绪标签之间的相关性识别方法，旨在从细粒度层级探索不同时长片段中二者之间的相关性。

（4）基于视觉交互行为的情绪识别研究

前几章的研究内容围绕虚拟交互环境中连续情绪测量及其与视觉交互行为、生理信号之间的相关性识别，没有面向行为特征构建用户独立的连续情绪Ground-Truth标签。为了解决这一问题，本文提出基于视觉交互行为的情绪识别方法。针对用户个体之间的认知差异与反应延迟，研究建立基于诱发素材参考特征的实时连续情绪标注序列时间对齐方法；针对虚拟体验中用户视觉交互行为的自由性与多样性，研究根据视觉行为模式进行片段层级视口聚类，提出一种视口相关的情绪融合方法，旨在细粒度层级上理解并分析虚拟环境中用户的情绪状态与对应诱发情境之间的关联性。

1.5 论文组织结构

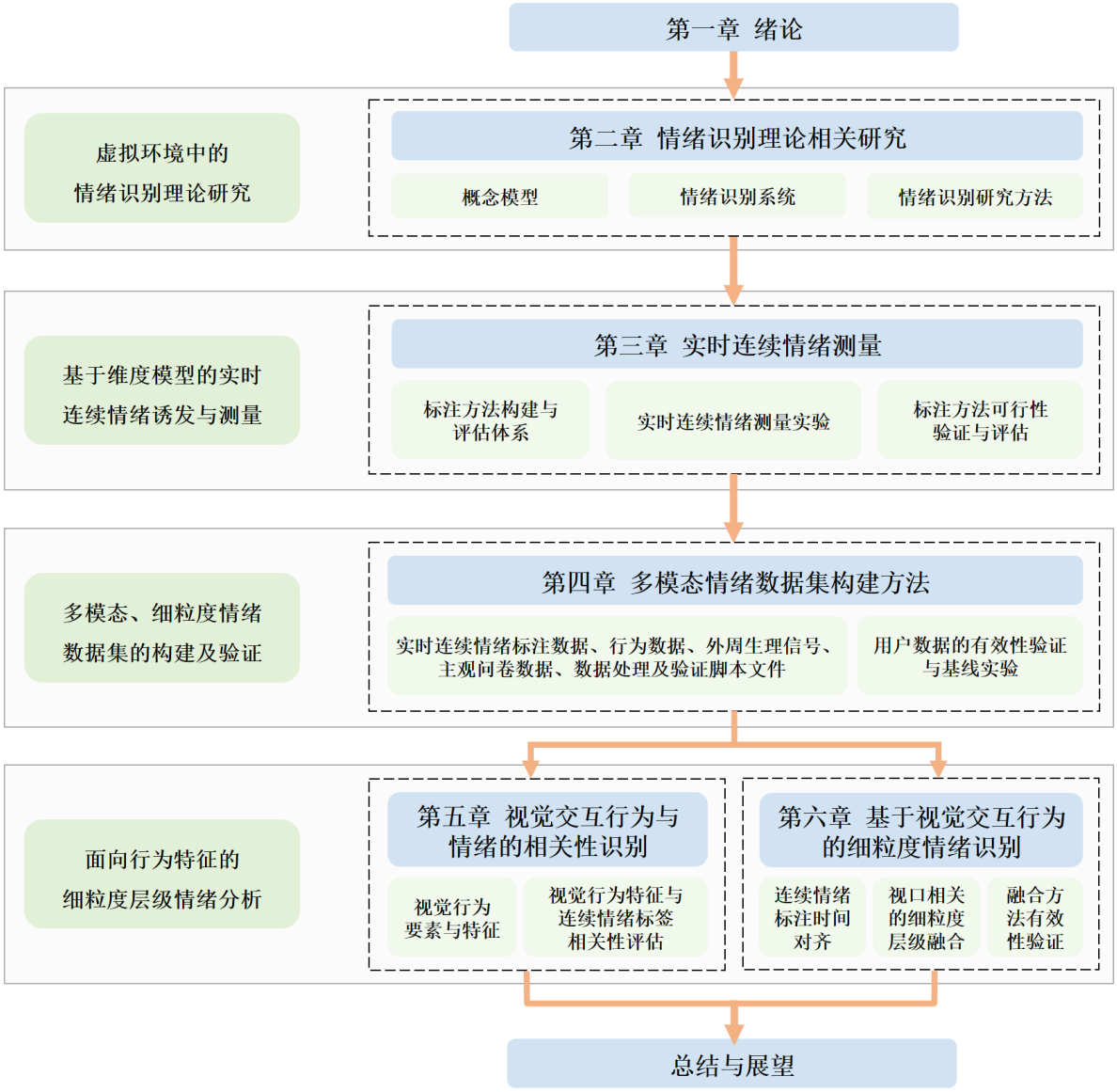


图 1.3 本文各章研究内容及之间的关系

本文各章节内容及相互之间的关系如图1.3所示，具体安排如下：

第一章阐述了本文的研究背景及学术与应用价值；介绍了虚拟环境中的交互行为及情绪识别研究现状，并分析了当前存在的问题；最后，介绍了本文各章节的研究内容及其内在关系和逻辑。

第二章开展了虚拟环境中的情绪识别理论研究。首先，提出了情绪识别概念模型；然后，从四个模块分析了情绪识别系统；最后，围绕多模态、细粒度情绪识别介绍了

相关研究方法。

第三章构建了虚拟交互环境中实时连续情绪标注方法与评估体系。首先，基于设计原则与共同设计环节形成两种标注方法；之后，建立了连续情绪标注方法评估体系；最后，构建了虚拟交互环境中实时连续情绪诱发及测量实验范式，并通过实验验证了基于 HaloLight 与 DotSize 的情绪诱发实验场景和情绪数据采集系统的可用性。

第四章创建了虚拟交互环境中首个公开可用的生理及行为情绪标注多模态数据集。首先，介绍了数据集结构、数据处理过程和存储格式；之后，对多模态用户数据进行预处理和统计学分析，验证了用户数据的有效性并介绍了数据集的多种使用方式；最后，通过一系列基线实验评估了数据集中生理与行为测量数据和实时连续情绪数据的关系。

第五章提出了视觉行为特征及其与连续情绪报告之间的相关性识别方法。首先，围绕头部运动与眼部运动探讨了用户的视觉行为特征；之后，提出了一种片段层级的用户头部运动、眼部运动与连续情绪标签之间相关性识别方法；最后，通过实验分析了虚拟体验中的用户视觉行为、以及不同时长片段中视觉行为与连续情绪标签之间的相关性。

第六章提出了基于视觉交互行为的情绪识别方法。首先建立了基于诱发素材参考特征的实时连续情绪标注序列时间对齐方法；之后研究根据视觉行为模式进行片段层级用户视口聚类，提出一种视口相关的情绪融合方法；最后，通过实验在细粒度层级上分析验证该方法的有效性及合理性。

论文最后对全文进行了总结，探讨了本文提出的方法和数据集的创新性和应用价值，并对未来的研究工作进行展望。

第 2 章　情绪识别理论相关研究

2.1　引言

人类对情绪的探索与研究由来已久，情绪认知理论（Cognitive Theory of Emotion）[14] 认为，情绪的产生是环境事件、生理状况和认知过程三种因素作用的结果。可见，情绪经由情境诱发产生，其类型、强度等受到人自身因素影响；同时影响着人的生理、心理及行为状态。积极的情绪有助于提升人的幸福感与工作效率，而消极的情绪可能会导致健康问题。情绪识别借助计算机相关技术理解人类情绪，提升技术与人的交互方式，更好地服务于人类。从理论角度来讲，情绪识别是一种基于生理、行为等情绪相关信号的模式识别；从实践角度来讲，情绪识别是情感计算的基石，旨在研发能够诱发、监测、理解人类情绪状态的方法工具。由于情绪会引起复杂的生理、心理及行为交互作用，即时、精准有效地识别人类情绪仍然是学术与各应用行业的一项挑战。

本章基于情感计算领域的已有工作，面向虚拟交互环境在理论层面上开展情绪识别研究。首先构建了人的情绪识别概念模型，从虚拟空间、情绪空间、模态空间及三者之间的逻辑关系进行问题的定义，并提出了情绪的可计算性、不确定性及同类情绪的相似性三种基本特性。然后，从情绪的建模、诱发、测量与理解四个环节阐述了情绪识别系统。最后，研究聚焦情绪的模态复杂性与时间复杂性，介绍了多模态、细粒度两种关键的情绪识别研究方法。

2.2　情绪识别概念模型

由于人类情绪具有主观性和复杂性，使其无法作为计算机直接处理的对象。为了借助计算机及相关技术实现情绪度量与计算模型，需要将其抽象为特定“问题”，进行形式化表示并作出基本假设和限定。本节提出虚拟环境中情绪识别的概念模型，如图2.1所示，研究介绍了相关模块的定义基本特性。

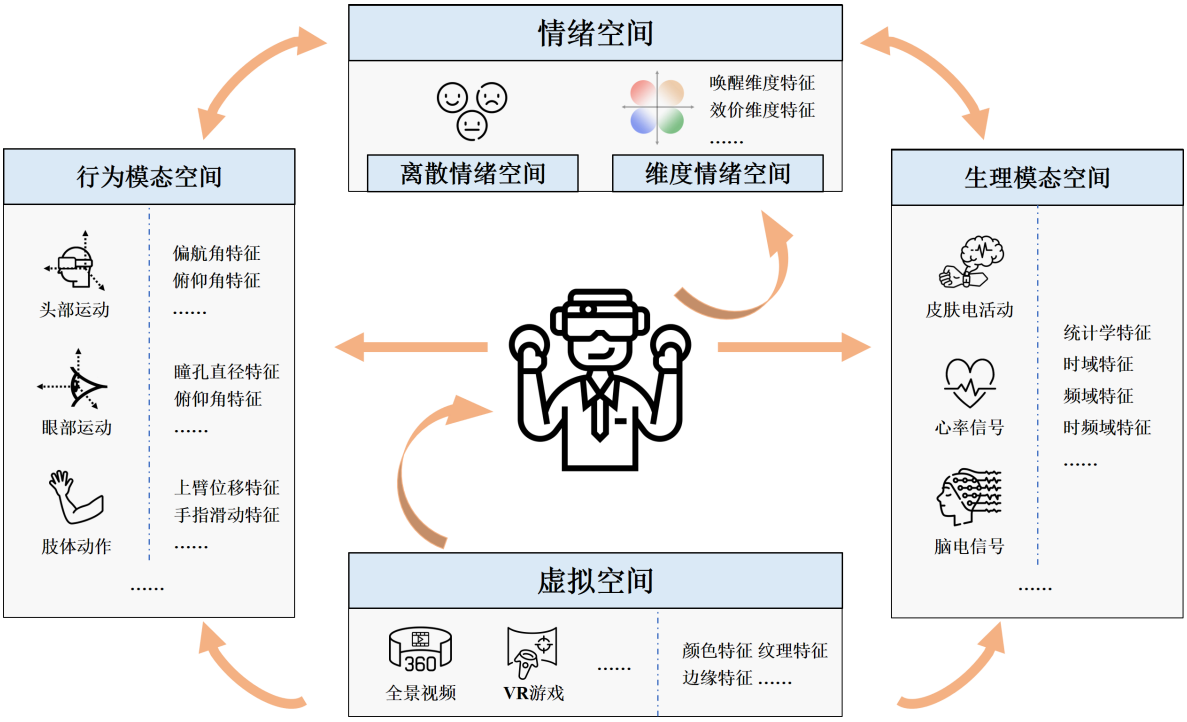


图 2.1　情绪识别概念模型

2.2.1　问题定义

定义 1（虚拟空间）

虚拟空间 VE 是指具体或抽象的全沉浸式三维环境，可以是真实环境的再现，也可以是完全计算机构建的人造世界。在虚拟空间中，人们佩戴 HMD 设备，通过头部运动控制视口变换，同时借助外部设备或身体动作等控制虚拟对象。可视化内容包括全景图片、全景视频、游戏任务等，用于情绪诱发的虚拟空间定义为：

$$V = \{V_1, V_2, \cdots, V_i, \cdots, V_m\}, \quad (i = 1, 2, \cdots, m) \tag{2.1}$$

其中 V_i 表示虚拟空间 V 的一个虚拟场景或时序性的全景视频等内容素材，m 表示素材的总数量。虚拟空间的内容素材中能够诱发情绪的每个细节或是属性称为虚拟空间的特征，从 V_i 中提取的内容相关特征构成 V_i 的特征空间，可定义为：

$$V_i = (v_{i1}, v_{i2}, \cdots, v_{ij}, \cdots, v_{in})^T, \quad (j = 1, 2, \cdots, n) \tag{2.2}$$

其中 v_{ij} 表示虚拟内容 V_i 的一种特征，n 表示提取的特征总数量。例如，包含三个全

景图片的虚拟空间可记为：$V=\{V_1,V_2,V_3\},m=3$；考虑全景图片 V_1 的颜色 $Color$、纹理 $Texture$、边缘 $Edge$ 三种视觉特征，则 $V_1=(Color,Texture,Edge)^T,n=3$。

定义 2（情绪空间）

通过视觉、听觉、嗅觉、味觉、触觉等多通道、多维度刺激人的大脑皮层，能够引发不同程度和不同类型的情绪状态。在虚拟环境中，人的视听通道均由虚拟空间所覆盖，由此诱发的情绪状态构成人在虚拟环境中的情绪空间 $Emotion$。情绪通常采用离散情绪模型和维度情绪模型两种表示方法。

定义 2.1（离散情绪空间）

虚拟环境中人的情绪可表示为若干种基本情绪状态，复杂的情绪状态可以通过基本情绪状态组合生成，所有基本情绪状态构成的集合称为离散情绪空间 $E_{category}$，可定义为：

$$E_{category}=\{E_1,E_2,\cdots,E_i,\cdots,E_m\},\ \ E_i=i(i=1,2,\cdots,m) \tag{2.3}$$

其中 E_i 表示一种基本情绪类型，m 表示基本情绪状态的总数量。例如可以设定 $E=\{1,2,3,4\},m=4$（1= 高兴, 2= 愤怒, 3= 恐惧, 4= 悲伤）。$E_{category}$ 中的任意两种基本情绪状态相互排斥，且任意两种情绪状态之间可以相互转移。

定义 2.2（维度情绪空间）

情绪相关的每个因子称为维度，其量化的数据类型有非数值型和数值型两种；非数值型如“高、中、低”等需要转化为相应的数值特征。情绪相关的所有维度构成了维度情绪空间 $E_{dimension}$，可定义为：

$$E_{dimension}=(X_1,X_2,\cdots,X_i,\cdots,X_n)^T,\ \ (i=1,2,\cdots,n) \tag{2.4}$$

其中 X_i 表示一种情绪维度，n 表示情绪维度总数量。情绪空间 $E_{dimension}$ 是一个 n 维列向量，即：$\mathbb{R}^n$；其中的任意两个情绪维度之间相互独立，每个采样点的情绪数据是这个情绪维度空间中的一个点。例如，Russell 提出的著名的“唤醒-效价（Arousal-Valence）”二维情绪模型[24]，$E=(Valence,Arousal)^T,n=2$；其中，效价维度表示情绪状态的愉悦程度，唤醒维度（又称激活）表示情绪状态的兴奋程度。

定义 3（模态空间）

与情绪相关的面部表情、肢体动作、心率血压等可测量指标统称为模态，各种可

测量指标构成了模态空间 $Modality$。在虚拟环境中，人的行为模态和生理模态是最常使用的情绪识别相关模态。从模态中提取的相关因子或属性称为特征，模态的各种特征矢量构成了模态特征空间。

定义 3.1（行为模态空间）

虚拟环境中的交互行为主要有视觉交互行为、肢体动作交互行为和外部设备交互行为。与情绪相关的行为模态空间可定义为：

$$M_{behavior} = \{B_1, B_2, \cdots, B_i, \cdots, B_m\}, \quad (i = 1, 2, \cdots, m) \tag{2.5}$$

其中 B_i 表示一种行为模态，m 表示情绪相关的行为模态总数量。从 B_i 中提取的行为特征构成模态 B_i 的特征空间，可定义为：

$$B_i = (b_{i1}, b_{i2}, \cdots, b_{ij}, \cdots, b_{in})^T, \quad (j = 1, 2, \cdots, n) \tag{2.6}$$

其中 b_{ij} 表示 B_i 模态的一种特征，n 表示提取的特征总数量。例如，考虑虚拟空间中最重要的视觉行为模态，包括头部运动、眼部运动和瞳孔直径，分别记为：HM、EM、PD，则 $M_{behavior} = \{HM, EM, PD\}, m = 3$；考虑头部运动相关的俯仰角均值 $Pitch_{mean}$、俯仰角中值 $Pitch_{median}$、偏航角均值 Yaw_{mean}、偏航角标准差 Yaw_{std} 四个特征值，则 $HM = (Pitch_{mean}, Pitch_{median}, Yaw_{mean}, Yaw_{std})^T, n = 4$。

定义 3.2（生理模态空间）

近年来，生理心理学家挖掘出不同的情绪状态及生理关联中的很多规律，情绪相关的生理模态空间可定义为：

$$M_{physio} = \{P_1, P_2, \cdots, P_i, \cdots, P_m\}, \quad (i = 1, 2, \cdots, m) \tag{2.7}$$

其中 P_i 表示一种生理信号模态，m 表示情绪相关的生理模态总数量。从 P_i 中提取的生理信号特征构成模态 P_i 的特征空间，可定义为：

$$P_i = (p_{i1}, p_{i2}, \cdots, p_{ij}, \cdots, p_{in})^T, \quad (j = 1, 2, \cdots, n) \tag{2.8}$$

其中 p_{ij} 表示 P_i 模态的一种特征，n 表示提取的特征总数量。例如，考虑虚拟空间中常用的皮肤电活动、心率变异性、脑电信号，分别记为：EDA、HRV、EEG，则 $M_{physio} =$

$\{EDA, HRV, EEG\}, m = 3$；考虑心率变异性的均值 HRV_{mean}、中值 HRV_{median}、标准差 HRV_{std} 三个特征值，则 $HRV = (HRV_{mean}, HRV_{median}, HRV_{std})^T, n = 3$。

定义 4（情绪诱发）

人的情绪是其所处的虚拟环境 V 与个体因素 H 相互作用的结果，可定义为：

$$E = \Omega(V \cdot H) \tag{2.9}$$

其中 Ω 表示函数关系。在虚拟环境中，人的视听通道均由虚拟内容所覆盖，其中的任何细节都能够诱发用户广泛的情绪状态。受到文化背景、性格特征、认知水平等影响，不同的人针对相同的诱发情境会产生不同的情绪反应；即使情绪状态一致，个体所表现的生理与行为表征也可能存在差异。

定义 5（情绪识别）

虚拟环境中的情绪识别是指精准有效地定义、测量并理解人在虚拟交互体验中的情绪状态，探索情绪空间和行为、生理等多模态空间之间的映射关系，可定义为：

$$E = f_1(M_{behavior}), \quad E = f_2(M_{physio}), \quad E = f_3(M_{behavior} \cdot M_{physio}) \tag{2.10}$$

其中 f_1、f_2、f_3 表示函数关系。由此，情绪识别是一个高度交叉的领域，有关信号处理、机器学习、生理心理学等。

2.2.2 基本特性

（1）情绪的可计算性

从一般意义来讲，可计算性是指计算机是否可以解决某一类实际问题。可计算性是情绪识别研究的基本前提，对可计算性的理解决定了情绪识别的研究思路和研究方法。情绪的定量研究是基于理性与逻辑推理，采用数学工具而实现的研究方法。传统的情绪定量研究主要依赖于问卷量表、结构化访谈等，计算机视觉、可穿戴式设备以及生理心理学相关技术扩展了情绪定量数据的收集范围。将行为、生理等可测量模态作为情绪数据的增益补充，使得情绪识别研究更为系统客观。情感计算研究认为情绪空间 E 中的所有样本点在模态空间 M 中都有对应的样本存在，对应关系包括以下四种：一对一关系、一对多关系、多对一关系、多对多关系。如果将情绪空间 E 作为自变量，行为、生理等模态空间 M 作为因变量，则条件概率 $P(M|E)$ 表示某种情绪状

态产生某种客观模态的概率；相反，可以用条件概率 $P(E|M)$ 表示某种模态指示或推理某种情绪状态的概率。

（2）情绪的不确定性

情绪依赖于虚拟环境的实时性变化,因此情绪信号具有不确定性;从数理统计的角度来看，情绪是维度空间中的随机向量，同类型情绪的不同样本是情绪空间的观测样本，通过对情绪空间分布规律的估计可以引导情绪分类。假设 $E=(X_1,X_1,\cdots,X_n)^T$ 是一个 n 维情绪向量,E 的所有可能取值为实数域的随机向量 $x,x=(x_1,x_2,\cdots,x_n)\in\mathbb{R}^n$，则情绪 E 的（联合）概率密度函数定义为：

$$f(x_1,x_2,\cdots,x_n) \tag{2.11}$$

（联合）概率分布函数为：

$$F(X_1,X_1,\cdots,X_n)=\int_{-\infty}^{x_1}\int_{-\infty}^{x_2}\cdots\int_{-\infty}^{x_n}f(x_1,x_2,\cdots,x_n)\,dx_1dx_2\cdots dx_n \tag{2.12}$$

（3）同类情绪的相似性

属于同一类别的情绪是由于对应模态的特征或是情绪本身的标签是相似的，情绪分类就是根据特征或标签之间的相似程度进行划分。例如根据模态 B_i 的特征向量 $(b_{i1},b_{i2},\cdots,b_{ij},\cdots,b_{in})^T$，将其归入 c 个类（$\omega_1,\omega_1,\cdots,\omega_c$）中。本文在4.5节开展了基于行为特征及生理信号的情绪分类实验。描述两个情绪样本之间的相似性主要有两种方法：相似系数和距离函数。相似系数的值可以刻画样本之间的相似性，样本点相似系数值越接近 1，相似性越高；相似系数值越接近 0 相似性越低。距离函数是较为常用相似性度量指标。假设 x_i 与 x_j 是两个情绪样本数据，二者之间的相似性度量 $\delta(x_i,x_j)$ 满足：非负性（$\delta(x_i,x_j)\geqslant 0$）和对称性（$\delta(x_i,x_j)=\delta(x_j,x_i)$）。两个情绪样本在 D 维欧几里得空间中的欧式距离为：

$$\delta(x_i,x_j) = \sqrt{\sum_{k=1}^{D}(x_{in}-x_{jn})^2} \tag{2.13}$$

2.3 情绪识别系统

虚拟环境中的情绪识别系统是由情绪建模、情绪诱发、情绪测量、情绪理解四个环节组成，如图2.2所示。

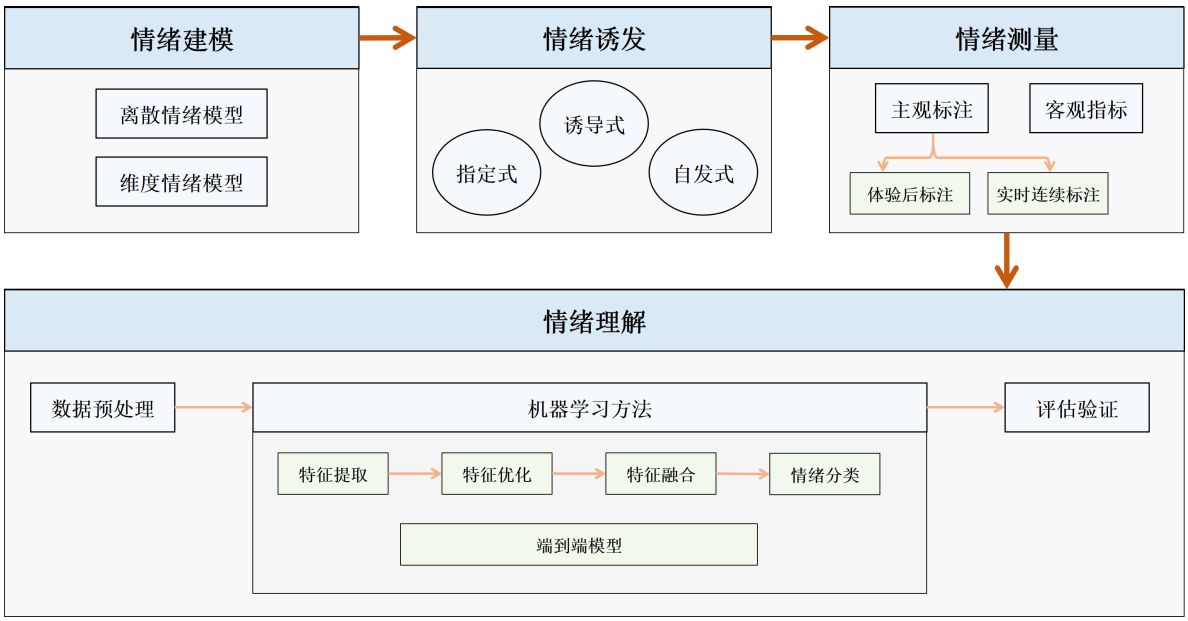


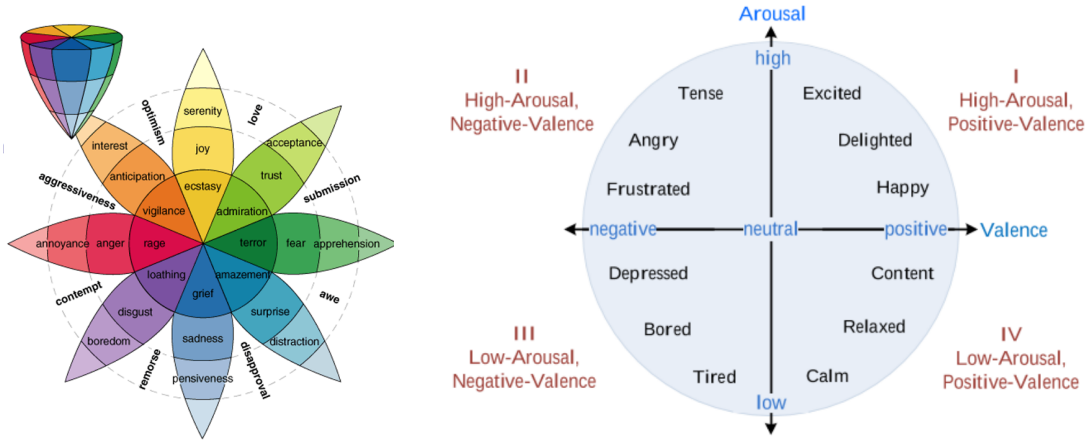
图 2.2 情绪识别系统示意图

2.3.1 情绪建模

情绪建模是情绪识别工作的第一步。近一个世纪以来，寻找一种理解与描述情绪的有效方法一直是学术界探讨的主题。离散情绪模型采用开心、恐惧、悲伤等离散标签描述情绪，Ekman 等人[85] 基于 Darwin 的研究[86] 和多项实验，总结出六种基本情绪标签，即：E_{Ekman} ={高兴，悲伤，生气，害怕，惊讶，厌恶}，$m = 6$。Plutchik[23] 进一步提出了由八种基本情绪组成的轮子情绪模型，即：$E_{Plutchik}$ = {开心，信任，害怕，惊讶，悲伤，厌恶，生气，期望}，$m = 8$，如图2.3(a)所示。离散情绪模型使用描述性词语表示情绪，具有易于理解、复杂度低等优势，经常在情绪研究中使用；但是，该类模型无法定量分析情绪状态和精确描述复杂情绪。

通过引入连续变量值，维度情绪模型可以在更精细粒度上描述更广泛的情绪状态。例如 Russell[24] 提出的 Circumplex “唤醒-效价” 二维情绪模型，即：$E_{Russell} = (V, A)^T, V, A \in [-1, 1]$，如图2.3(b)所示；尽管该模型在情绪研究中广泛使用，但无法区分在效价和唤醒维度上具有相似值的情绪状态，如 “恐惧” 与 “愤怒”；Mehrabian[87]

引入了支配（Dominance）维度（又称控制），用于表示人对当前情境的控制程度；提出“效价-唤醒-支配”三维情绪模型（Valence-Arousal-Dominance, VAD），即：$E_{Mehrabian} = (V, A, D)^T$；“愤怒”在控制维度的值较高，而“担心”在该维度值较低。Fontaine 等人[88] 提出引入期望（Expectation）作为第四维度，即：$E_{Fontaine} = (V, A, D, E)^T$。该四维模型可以从其他情绪状态中区分“惊讶”，但仍然难以区分“羞愧”、“内疚”和“尴尬”三种状态。维度情绪模型的另一个问题在于，由于人的认知与反应需要一段时间，持续性输入过程中所记录的情绪状态和真实情绪状态之间存在时间延迟。上述所有模型中，“唤醒-效价”二维情绪模型在情绪识别研究中的使用率最高，研究[89] 表明该模型可以描述绝大多数常见的情绪状态，且机器学习识别算法模型构建的复杂度低、识别结果良好。



(a) 轮子情绪模型[23]　　(b) “唤醒-效价”情绪模型[24,90]

图 2.3　离散与连续两种情绪表示模型

2.3.2　情绪诱发

情绪识别的第二步是情绪诱发。正确的诱发机制对于研究质量和结果的可靠性至关重要。情感计算研究[91] 中主要有三种情绪诱发方式：（1）指定式，通过给用户提出要求指定情绪状态，如快乐或悲伤；该方法最容易诱发并捕捉情绪状态，但研究[16] 表明在这一方式中，人们往往会夸大情绪表现。（2）诱导式，将用户置于特定环境中诱发情绪状态，常用的诱发素材包括音乐、图片、视频、场景等；该方法能够产生更自然的情绪反应，但受限于诱发素材，无法诱发所有复杂的情绪状态[92]，仍然是一种被动诱发方式。（3）自发式，模拟真实环境中日常生活情境，如面对面沟通、电话交

谈、辩论等；该方法更贴合人在真实环境下的自然状态，但由于情绪状态的瞬时性、上下文相关性，很难捕捉用户准确的情绪状态。

虚拟环境具有沉浸感与构想性，为用户提供了视觉、肢体动作、外部辅助设备等丰富的交互模式，带来沉浸式交互体验。诱发素材包括全景图片[35]、全景视频[63]、基于交互场景的任务与游戏[29] 等。相比于非沉浸环境和半沉浸环境，全沉浸式虚拟环境的临场感和交互性能够让用户更深入地融入到诱发素材中，在与之交互中更可靠的诱发人的多样情绪状态。另一方面，伴随着 HMD 设备的多样化和迅速商业化，虚拟环境更容易搭建并部署服务于终端用户。因此，虚拟环境能够在可控的环境中诱发用户更广泛的情绪状态，是情绪识别的强有力工具。

2.3.3 情绪测量

情绪测量是情绪识别研究的必要先决条件和原始依据。人的情绪作为一种复杂的生理心理现象，测量主要有主观标注和客观指标两种方法。

主观标注是最直观的情绪测量方式，由用户直接给出情绪状态自我报告（Self-Report），包括体验后标注和实时连续标注两类。体验后标注方法有基于李克特量表（Likert）的调查问卷、小组访谈等；例如图片导向型自我评估工具（Self-Assessment Manikin，SAM）[93]，将情绪划分为若干离散类别，采用不同的表情图片作为量表，从唤醒和效价（和支配）维度获取用户的情绪报告。该类方法忽略了时序性诱发素材中情绪的连续本质，会产生情景记忆偏差等问题。实时连续标注是指体验过程中实时报告情绪状态，一些研究[78,94,95] 提出了基于手持设备的实时连续情绪标注方法；用户通过控制器等设备基于维度情绪模型给出连续情绪评级，同时能够获得标注结果的实时反馈。Metallinou 等人[20] 探讨了情绪连续标注方法的机遇与挑战，指出连续情绪标注任务中标注员自身文化背景、情绪属性定义、标注设备可用性、诱发素材先验知识、标注员反应延迟五个难点，强调了从标注员个体之间一致性、标注员重复标注一致性、实时标注与离散标签之间关系三个方面分析标注数据的重要意义。

客观指标是指用户所表现出来的可测量状态，包括生理信号、行为动作、面部表情等多种模态指标。行为模态用于情绪测量的优势在于人体有较多可以被监测的参考点，通过计算机视觉方法或是穿戴式动作捕捉设备能够跟踪这些参考点的运动，但是微表情等不明显动作难以监测分辨，且人可以控制自身行为动作。相比于此，人们很难控制生理反应，因此这些信号可用于揭示隐藏的情绪状态。情绪相关的生理信号

主要来自中枢神经系统与外周神经系统[89]，包括 EDA、原发性皮肤电反应（Galvanic Skin Reponse，GSR）、HRV、EEG、肌电图（Electromyography，EMG）等。但是在运动过程中测量生理信号，信号的稳定性仍是一项挑战[16]。

虚拟环境中现有的生理信号和行为监测方式均具有较大的局限性；目前也尚无科学有效的针对虚拟交互环境的实时连续情绪标注方法和工具。

2.3.4 情绪理解

情绪理解是情绪识别最为重要的一步。该环节完成的是建立多模态客观指标和主观标注之间的映射关系。采集自用户个体的多模态数据特别是生理信号非常复杂，带有主观性且对串扰、多传感器、运动伪影等干扰具有敏感性，首先需要对这些原始数据进行滤波、降噪、时间同步等预处理。

传统的统计学方法是使用最广泛的情绪理解方法，用于研究用户情绪状态和生理、行为等可测量指标之间相关性及显著性。特别是针对虚拟环境，Marín-Morales 等人[96] 指出现有的情绪识别研究中 88% 的学术论文选择了统计学方法。最常使用的统计学方法主要为以下三种：（1）推断统计分析，根据样本数据推断总体特征，主要包括总体参数估计与假设检验两个部分；（2）相关性分析，针对客观事物之间的统计关系进行定量分析，包括 Pearson 相关系数、Spearman 相关系数、Kendall’s tau -b 相关系数；（3）聚类分析，将一批数据按照选定特征的亲疏程度进行分类，类内部个体特征具有相似性、类间个体特征的差异性较大，主要包括层次聚类和 K-均值聚类两种。该类方法简单直观，在一定程度上能够准确、快速获取重要且有效的信息。当前，传统的统计学方法存在以下三个缺点[96]：（1）基于均值和偏差的假设检验方法能够分析两组样本之间的差异，但无法进行更精细程度的情绪识别；（2）对于包含较多变量的数据集，很难分析多种特征组合的识别效果；（3）没有考虑多模态数据结构中的非线性关系。

机器学习方法可以自动识别复杂数据模式，需要精心设计并选取最显著的模态时域、频域特征，在完成特征选择或是冗余特征降维等优化处理后进行特征融合，优化与融合之后的特征序列用于情绪分类。常用于情绪理解的机器学习分类模型有 SVM[97]、随机森林 RF[98]、高斯朴素贝叶斯（Gaussian Naive Bayes，GaussianNB）[99] 和 K-最近邻（K-Nearest Neighbor，K-NN）[100] 等。相比于这些传统的机器学习方法，基于神经网络（Neural Network，NN）的深度学习方法进行情绪分类，能够学习数据固有的分布、

并自动提取特征，是一种端到端的模型；常见的情绪识别深度学习算法有卷积神经网络（Convolutional Neural Network，CNN）、长短记忆模型（Long Short-Term Memory，LSTM）[101]、循环神经网络（Recurrent neural network，RNN）[102] 等。但是，深度学习模型结构比较复杂，训练模型需要大量数据；有限的情绪识别数据样本量，在一定程度上制约了端到端情绪识别模型的发展。为了评估机器学习方法中情绪模型准确度，采用用户依赖（Subject-Dependent，SD）和用户独立（Subject-Independent，SI）两种验证方法，其中 SI 验证方法可以更好地评估模型泛化能力。

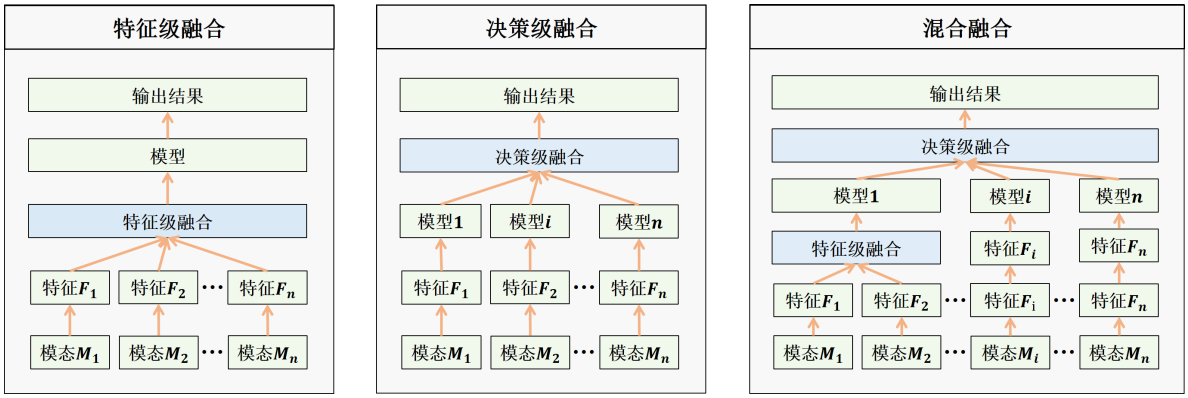
2.4 情绪识别研究方法

本节聚焦情绪的模态复杂性和时间复杂性，从多模态情绪识别和细粒度情绪识别两个方面介绍情绪识别研究方法。

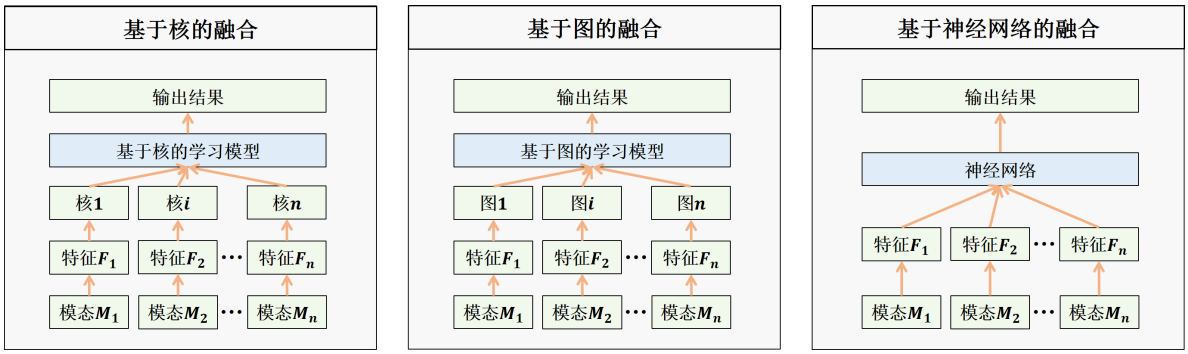
2.4.1 多模态情绪识别

语音、姿态、生理信号等单一模态的情绪识别研究由来已久，并伴随着深度学习的进步发展迅速。但是，人的情绪表达是多元化的，单模态情绪识别难以准确判断人的情绪状态。例如语音、表情等单模态数据受限于主观因素，并不总能观察到明显准确的外部表现；生理信号虽然难以隐藏，但测量难度大，其稳定性易受运动干扰。相比于此，多模态情绪识别整合了模态之间的互补信息，可以构建更健壮的情绪识别体系。荟萃分析[103] 表明多模态情绪识别比最好的单模态识别的平均性能提高了 9.83%。但是，不同模态的情绪表达方式差异大，不同模态的特征数据提取、分析与识别方法也不相同；如何将不同模态的特征融合处理，充分挖掘模态间的特征信息并去除冗余，是多模态情绪识别的核心问题。多模态特征融合主要有模型无关融合（Model-Free Fusion）和基于模型融合（Model-based Fusion）两种方法，如图2.4所示。

模型无关的融合不依赖于特定学习算法，包括特征级（Feature-Level Fusion）融合、决策级（Decision-Level Fusion）融合和混合（Hybrid Fusion）融合三种方式，如图2.4(a)所示。（1）特征级融合利用每个模态特征之间的相关性与相互作用，将多模态特征空间串联形成一个特征空间，仅需要单一模型训练；该方法非常直观，且最大限度保留了模态空间中的原始信息，但是在融合前需要将不同模态特征进行时间同步。（2）决策级融合采用独立模型训练每个模态，再将多个模型的输出结果通过取均值、投票法、集成学习等机制进行融合；该方法更为灵活，可以为不同模态选择最佳



(a) 三种模型无关的情绪融合方法，其中 n 表示模态总数量



(b) 三种基于模型的情绪融合方法，其中 n 表示模态总数量

图 2.4　多模态情绪融合方法

分类器，且识别系统能够随着模态数的变化动态修改，但忽略了各模态在特征层面的相关性。（3）混合融合结合了上述两种融合方式的优势，但同时也带来了更大的计算成本。

基于模型的融合在学习模型构造过程中进行特征融合，主要包括基于核的融合（Kernel-based Fusion）、基于图的融合（Graph-based Fusion）和基于神经网络（Neural Networks）的融合三种方式，如图2.4(b)所示。（1）基于核的融合是核支持向量机（Kernel SVM）的扩展，对于不同的数据模态采用不同的核；核选择的灵活性与损失函数的凸面性使得多核学习融合在多模态情绪识别领域很受欢迎，其缺点在于测试期间依赖大量的训练数据，使得推理速度慢且造成大量的内存消耗。（2）基于图的融合为每个模态构造独立图或超图，将这些图组合成为一个融合图并通过基于图的学习计算不同边和模态的权重；该方法能够很好地处理数据不完整性问题，还可以在对应边中融入先验知识，但计算成本也会随着训练样本的增多而成倍增加。（3）基于神经网络的融合是将所有可用模态的信息直接与深度学习网络相结合，如基于注意力的融合

（Attention-based）；该方法能够从大量数据中学习，可以将多模态信息深度融合，提高识别性能，但神经网络的可解释性差且需要大量的训练数据。

由于人的情绪反应存在较大的个体差异，且情绪相关模态测量难度大、成本高，现有的多模态情绪研究主要关注音频、视频和文本模态的结合，而与人相关的生理信号、行为等多模态情绪识别研究相对较少。Liu 等人[104] 基于 SEED 与 DEAP 数据集，采用受限玻尔兹曼机（Restricted Boltzmann Machine，RBM）融合 EEG 与眼部运动进行双模态情绪识别。Wu 等人[105] 提出一个带有注意力机制的层次 LSTM 模型，用于融合面部特征和 EEG 特征计算情绪状态。虚拟环境中现有的多模态情绪识别主要采用假设检验、相关性分析等统计学方法分析生理信号与情绪状态之间的对应关系；由于缺乏情绪相关的行为、生理数据及情绪 Ground-Truth 数据，虚拟环境中的多模态情绪识别仍处于起步阶段。

2.4.2 细粒度情绪识别

相比于单个图片，视频、游戏等诱发素材具有时序性信息，近年来的情绪识别研究开始关注情绪的时间动态建模。不同于在一段体验中识别一种情绪，细粒度情绪识别（Fine-grained Emotion Recognition）[42] 指的是模型在一个特定时间间隔内借助生理、行为等情绪相关信号输出多种情绪状态。为了捕获并识别人类情绪的时序性变化，一方面需要考虑如何将来自多个标注员的实时连续标注序列信息融合，获取对于特定诱发内容用户独立的情绪 Ground-Truth 标签。另一方面，研究需要使用或开发实时连续的情绪识别算法与模型。

尽管目前已经出现了很多情绪识别算法，但细粒度情绪识别仍处于起步阶段[106]。细粒度情绪识别通常采用回归和分类两种方法：（1）在回归方法中，研究将目标情绪状态视为一个连续序列，直接计算从输入信号到输出情绪序列的映射，包括 LSTM[101]、支持向量回归（Support Vector Regression，SVR）[107]、多项式回归（Polynomial Regression，PR）[108] 等顺序学习方法；回归方法可以获得更高的识别准确度，但由于其循环结构的训练方式是从信号开始到信号结束，前序样本的回归误差积累会影响整个序列的结果。（2）在分类方法中，研究将连续情绪测量信号分割为不同时长的片段，并独立地对每个片段中的情绪进行分类；该方法中不同片段之间的识别结果不会相互影响，但主要挑战在于提取并融合片段内和片段间的特征值，以及片段内的信息是否足够用于识别情绪状态。最近，Romeo 等人[109] 首次将多示例学习（Multi-Instance Learning，

MIL）引入到基于生理信号的情绪识别研究中，采用基于 SVM 的 MIL 算法识别每个细粒度片段中的唤醒与效价类别，在“高”唤醒标签上准确率达到了 68%；Zhang 等人[110] 提出基于少样本学习（Few-Shot Learning）的 EmoDSN 算法，通过最大化多模态片段与情绪标签之间的距离矩阵实现细粒度情绪识别，效价与唤醒维度的二分类准确率为 76.04% 和 76.62%。

细粒度情绪识别离不开细粒度的情绪 Ground-Truth 标签，分类方法中 Ground-Truth 标签频率应与片段时长一致，通常为 5 秒或 10 秒等[111,112]；回归方法中的标签频率应与输入信号一致，例如生理或行为测量数据的采样频率。Romeo 等人[109] 曾指出，缺乏连续性标注数据是造成细粒度情绪识别中弱监督算法验证失败的原因。离散的情绪标签仅能反应用户在体验过程中最显著或是体验最后阶段的情绪状态，会造成机器学习算法过拟合问题；相比于此，实时连续情绪标签能够学习动态情绪变化与情绪相关模态信号之间更精确的映射关系，从而降低过拟合问题。尽管最近的研究试图构建离散或稀疏情绪标签（如体验后标注）用于情绪识别的模型框架，但仍然需要可靠的细粒度情绪标签进行验证与训练。由于缺乏实时连续情绪标注方法和多模态情绪相关测量信号，尚无针对虚拟交互环境中的细粒度情绪理解研究。

2.5 本章小结

本章基于情感计算领域的已有工作，阐述了虚拟交互环境中的情绪识别理论研究，主要包含三个方面的研究内容。第一方面是虚拟环境中的情绪识别概念模型，借助形式化语言表示方法，对虚拟空间、情绪空间、模态空间及三者之间的逻辑关系进行问题的定义，并提出了情绪的可计算性、不确定性及同类情绪的相似性三种基本特性。第二方面是情绪识别系统，虚拟环境中的情绪识别系统包含情绪建模、情绪诱发、情绪测量和情绪理解四个环节，研究指出了连续维度情绪模型能够在更精细粒度上描述更广泛的情绪状态；虚拟环境是情绪识别的强有力工具；由于客观测量方法的局限性，需要开发针对虚拟环境的实时连续情绪测量方法；相比于传统的统计学方法，机器学习技术能够更好地识别复杂情绪数据。第三方面是情绪识别研究方法，研究聚焦情绪相关模态的复杂性及实时连续情绪的时间复杂性，介绍了多模态情绪识别中模型无关与基于模型的两种融合方法、细粒度情绪识别中的分类与回归两种方法，以及相关的机器学习模型和算法；研究指出多模态、细粒度情绪数据的缺乏是虚拟环境中情绪识别亟需解决的问题。

第 3 章　实时连续情绪测量

3.1　引言

虚拟环境的沉浸感与交互性能够诱发用户广泛的情绪[113]。无论研究目的是监测、理解用户在教育教学[31]、军事训练[30]、医学诊疗[43]、游戏娱乐[29] 中的情绪，或是开发面向虚拟体验的情绪识别与自适应系统[35]，收集用户在体验过程中精确有效的情绪 Ground-Truth 标签非常重要[35]。传统的情绪自我报告发生在体验完成之后，用户根据研究内容从喜悦度、交互度、唤醒与效价维度等方面给出情绪评级[24]。这一方式本质上是回顾性和离散性的，因为用户在整个体验中会产生多种情绪[25,114]；特别是对于视频等时序性内容，例如用户在观看视频后将其标记为“开心”，但在整个体验过程中会经历不止一种情绪。此外，根据人机交互中关于用户自我报告的注解[115]，回顾性评估依赖于情景记忆，这可能会产生情景记忆偏差[115]，例如：峰终效应（Peak-End Rule）[116]。为此，一些研究转向实时连续情绪标注，开发并验证了面向电脑端的 FEELTRACE[94]、CASE[78] 及移动端的 RCEA[79] 等标注工具及方法。

在虚拟环境中，用户的视听认知通道均由 HMD 覆盖。Toet 等人[117] 指出，虚拟环境中现有的情绪自我报告方法耗时长、需要较大的认知负荷与理解；或是发生在虚拟环境之外，破坏了虚拟体验的沉浸感。目前尚无针对虚拟环境的实时连续情绪标注方法。本章采用人机交互的设计准则和研究方法，旨在获取用户在虚拟体验中准确有效的实时连续情绪报告，同时尽可能降低对临场感与沉浸感的干扰。主要研究内容为如下三个方面：

（1）提出了虚拟交互环境中实时连续情绪标注方法。研究聚焦标注方法的三个设计原则，选择高分辨率的 HMD 设备为用户提供高质量内容、采用带有摇杆的无线数字游戏控制器作为情绪标注设备，通过多领域专家共同设计（Co-Design）形成 HaloLight 和 DotSize 两种标注信息可视化方案。

（2）建立了虚拟交互环境中实时连续情绪标注方法评估框架。研究从用户体验质量和标注数据的有效性两个方面给出了评估指标及评估方法；用户体验质量采用显式和隐式两类方法分析晕动症、临场感和认知负荷三个指标，标注数据的有效性从唤醒和效价两个维度采用统计学方法进行评估。

（3）构建了虚拟交互环境中实时连续情绪诱发及测量实验范式。研究提出基于

HaloLight 和 DotSize 的情绪诱发实验场景和情绪数据采集系统，并采用隐式和显式两类测量法分析了用户的情绪标注数据与用户体验结果，针对实时连续情绪标注方法的有效性展开讨论。

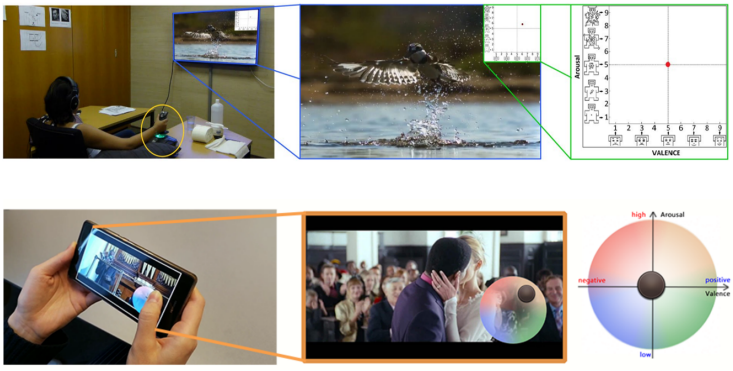
实验结果表明，HaloLight 与 DotSize 两种实时连续情绪标注方法能够收集虚拟交互环境中用户精确有效的唤醒与效价维度情绪 Ground-Truth 标签，两种方法在用户体验的晕动症、临场感和任务负荷方面没有显著性差异，且没有影响用户的虚拟体验质量。以下将会详细介绍研究方法和实验内容。

3.2 相关工作

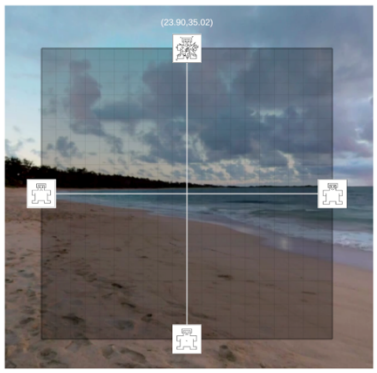
近年来，人机交互相关研究[16,118] 提出了实时连续的情绪标注技术，用以获取用户体验中细粒度的情绪标签；研究方法通常为，选择一个控制器作为输入设备，该设备可以提供连续的情绪评级，并将标注结果实时反馈给用户。考虑键盘-鼠标人机交互模式，相继出现了 FEELTRACE[94]、EmuJoy[114] 和 GTrace[119] 等标注工具。用户在二维空间中，通过连续点击鼠标标注情绪状态，但这一方式增加了用户的身体负荷和认知负荷[20,120]。一些研究将鼠标操作替换为更符合人体工程学的辅助设备：Girard 等人[121] 开发了一个一维滑条工具 CARMA，用户通过推动滑条上下滑动来报告情绪的消极或是积极状态；Lopes 等人[122] 研发的 RankTrace 工具是一个径向控制器，也可用于单一维度的连续情绪报告，例如情绪强度。为了实现唤醒-效价二维情绪空间中的标注，Girard 等人[95] 改进后的 DARMA 软件和 Sharma 等人[78] 开发的 CASE 工具引入了摇杆控制器。摇杆头的运动映射在唤醒-效价二维情绪空间中，摇杆运动的水平轴对应效价维度，竖直轴对应唤醒维度；用户的标注结果呈现在视频播放器旁侧的窗口中或是叠加在播放器的右上角，内容为二维坐标系和用户当前标注点所在位置。最近，Zhang 等人 RCEA[79] 设计了一种移动环境中实时连续情绪标注方法，用户在使用移动设备观看视频的同时，采用虚拟摇杆实时标注情绪状态，如图3.1(a)所示。

用户在虚拟体验中的情绪产生自虚拟环境，也应该在虚拟情境下报告[123]。Putze 等人[124] 指出，在虚拟环境中进行问卷调查不仅可以更便捷有效的获取情绪状态，也降低了临场感中断造成的情绪可信度降低等问题，避免了因环境转换而造成的数据偏差。最近，Toet 等人[117] 提出了 EmojiGrid 方法，在虚拟环境中内嵌一个笑脸网格用于情绪评估；Krüger 等人[125] 设计了更直观的表情变形（Moroh A Mood，MAM）方法，采用一套三维角色的面部表情作为情绪量表。这两种情绪自我报告方法虽然在虚

拟环境中进行，但是均发生在虚拟体验之后。Voigt-Antons 等人[126] 设计了一个固定的唤醒-效价二维情绪空间网格界面，叠加在虚拟环境中的全景视频之上，用户可以通过点击网格上的任意点对视频内容进行评估，如图3.1(b)所示。但是，作者并没有评估这一技术的可用性，也没有解决标注结果的融合问题；此外，固定在视口区域的网格图会严重影响用户观看体验，通过鼠标点击的方式也在很大程度上限制了用户与虚拟环境的自由交互。因此，需要开发更为便捷的基于无线控制器的输入设备用于虚拟环境中用户的实时连续情绪标注。



(a) 电脑端[78]（上）和移动端[79]（下）的情绪标注工具



(b) 虚拟环境中的情绪标注工具[126]

图 3.1　典型的实时连续情绪标注工具

在虚拟环境中，用户需要在进行虚拟场景交互、全景视频观看等首要任务的同时实时标注自己的情绪状态。因此，需要确保标注任务既没有增加用户的认知负荷、也没有分散用户在主任务上的注意力。人机交互中的外周视觉交互（Peripheral Visualization）研究[127,128] 表明，将信息显示在视觉注意核心区域的周边，能够帮助用户在进行主要任务的同时快速有效接受额外信息。在传统的电脑端视频观看任务中，Mairena 等人[129] 和 Gutwin 等人[130] 采用一系列视觉变量（颜色、形状和运动）为用户提供外周提示信息。在移动环境下，Zhang 等人[79] 考虑到移动设备的屏幕很小、且移动环境下用户容易分散注意力，采用外周视觉信息（边框、透明度、位置和大小等要素）来实时反馈用户的情绪标注状态。在虚拟环境中，Gruenefeld 等人[131] 提出了 HaloVR 和 WedgeVR 技术，采用外周线索引导用户的视觉注意力。本章也考虑采用外周可视化信息，为用户在虚拟体验中实时连续报告情绪状态提供标注反馈。

3.3 实时连续情绪标注方法

3.3.1 设计原则

针对虚拟交互环境中情绪标注不实时不连续、标注方法耗时长且干扰用户体验、细粒度情绪标注研究空白的问题，本节基于虚拟空间中人机交互的设计准则[132]，采用启发法（Heuristics）[133] 逐步缩小设计空间，聚焦以下三个设计原则：

（1）考虑基于 VR 头戴显示器设备的虚拟交互环境。研究[132] 表明，用户在虚拟环境中全程佩戴的 HMD 设备具有潜在的计算延迟、显示频闪、缺乏个性化校准、未满足人体工程学等问题，会带来感官冲突或姿态不稳定，从而让用户在虚拟体验中产生晕动症（Motion Sickness）[134]。长时间佩戴 HMD 设备也会带来视觉等疲劳感。由于实时连续情绪标注发生在虚拟交互环境中，因此需要考虑并尽可能避免这些问题。

（2）考虑情绪标注设备的人体工程学性能。用户在虚拟环境中因佩戴 HMD 设备而无法看到标注设备，因此需要确保情绪标注设备使用舒适、标注结果精准可信。相比传统的鼠标设备，摇杆控制器带有回位弹簧，摇杆可以在无作用力的情况下自动调整至中心位置，能够为用户提供本体感知反馈，这使得它更适用于在虚拟交互环境中进行连续标注[135]。

（3）考虑多重任务引起的注意力分散。在虚拟环境中，虚拟体验是用户的主要任务，额外的实时连续情绪标注任务会造成注意力分散[136,137]。因此需要降低标注任务带来的身体负荷与认知负荷，在不干扰用户体验的情况下实时提供标注状态反馈。基于人机交互的外周视觉信息研究，显示在用户视觉区域周边的信息能够帮助用户在进行主任务的同时快速有效地获取并理解信息[127]。借助图形用户界面里元素的尺寸、颜色、透明度[138,139] 等因素可以减少用户注意力分散，确保其对主任务的关注度。

考虑原则（1），选择高分辨率的 HMD 设备，为用户提供更高质量的内容。在虚拟环境中，HMD 设备完全覆盖用户视觉区域，内嵌有眼动测量功能的设备可以获取更多的数据维度。表3.1中列出了三种常见的内嵌有眼动测量设备的 HMD 属性参数。综合显示分辨率、视场角、刷新率、重量、开发难易度等方面，考虑 HTC VIVE Pro Eye HMD。该设备单个目镜的分辨率为 1440 × 1600，双目分辨率为 2880 × 1600，视场角为 110 度，刷新率为 90Hz；内置的眼动追踪仪双目注视数据输出频率为 120Hz，精度为 0.5 度。同时，选择时长较短的诱发素材，避免用户因长时间佩戴头显而造成的眩晕、疲劳等不适感。

表 3.1　常见的内嵌有眼动设备的头戴显示器设备参数表

设备名称	显示分辨率（像素）	视场角（度）	刷新率（赫兹）	质量（克）
HTC VIVE Pro Eye[1]	2880 × 1600	110	90	790
Pico Neo 2 Eye[2]	3840 × 2160	101	75	350
StarVR One[3]	1,600 万次	210 水平 × 130 垂直	90	450

[1] https://enterprise.vive.com/us/product/vive-pro-eye/
[2] https://www.pico-interactive.com/us/neo2.html
[3] https://www.starvr.com/products/

考虑原则（2），选择带有摇杆的无线数字游戏控制器 Joy-Con[1]作为标注设备。为了提升操作灵活度，在摇杆头处增加一个 11 毫米高的帽子，以延长摇杆长度。基于“唤醒-效价”二维情绪空间[24]，摇杆在水平方向上的运动表示情绪的效价维度值，竖直方向上的运动表示情绪的唤醒维度值，横纵坐标形成的四个象限分别表示四种类型的情绪信息，如图3.2所示。第一象限表示情绪积极且开心（High Arousal High Valence，HAHV），第二象限为积极但不开心（High Arousal Low Valence，HALV），第三象限为不积极且不开心（Low Arousal Low Valence，LALV），第四象限为积极但不开心（Low Arousal High Valence，LAHV）。标注过程中摇杆头与摇杆中心的距离越远，表示情绪强度越大。

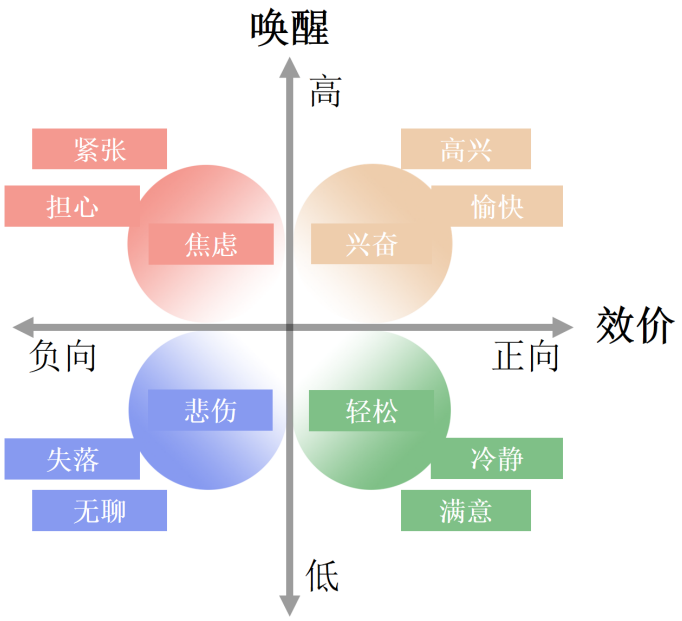
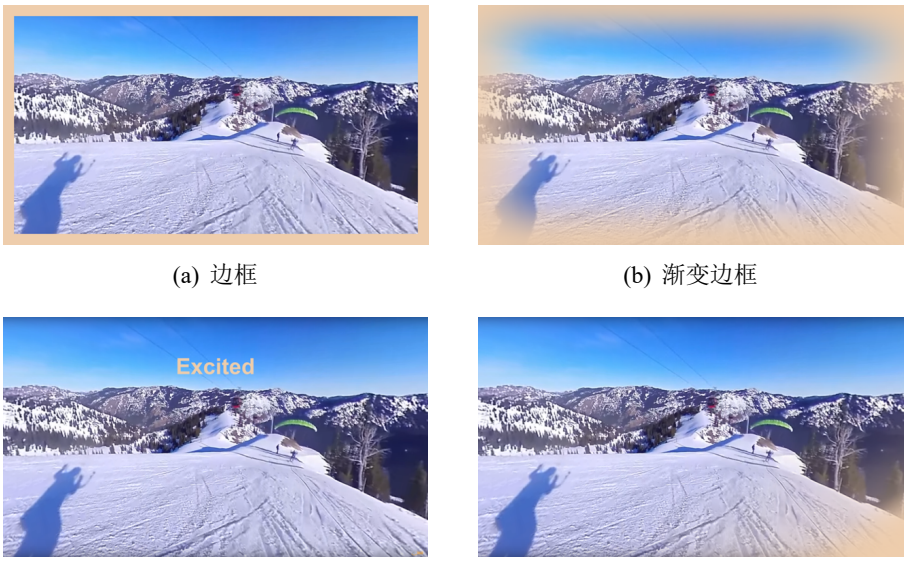


图 3.2　“唤醒-效价”二维情绪模型

[1]https://www.nintendo.com/switch/choose-your-joy-con-color/

考虑原则（3），研究基于边框、实心圆、灯光、文字四种界面设计元素的位置、大小与透明度等属性，提出以下四种标注信息外周可视化方案：边框（图3.3(a)），包围在视口外周的颜色边框；渐变边框（图3.3(b)），包围在视口外周的渐变颜色边框；文字（图3.3(c)），呈现在视口中心偏上位置的文本标签；光束（图3.3(d)），呈现在视口右下角的渐变颜色光束。上述方案里，元素颜色表示用户当前正在标注的情绪类型，颜色的选择基于简化版的 Itten 颜色系统[140]，先前工作表明该系统是直观且易于理解的[141]。情绪模型中的第一象限至第四象限，颜色的十六进制值依次为 #eecdac、#f4978e、#879af0、#7fc087，如图3.2所示。



(a) 边框　(b) 渐变边框

(c) 文字　(d) 光束

图 3.3　基于外周视觉信息的四种原型设计方案

3.3.2　HaloLight 与 DotSize

本节采用以用户为中心的方法[142]，在专家共同设计（Co-Design）[143] 的基础上，对图3.3中四种外周视觉信息的标注原型方案进行多轮迭代评估研究。这一环节共有六名研究员参与（P_i 为参与者 $i \in [1,6]$），年龄区间为 25-39 岁（$M = 32.8, SD = 4.8$），均有一年以上的 VR 研究经历。其中三位是人机交互研究员，一位是交互设计师，两位是软件工程师。本环节的流程如下：首先，向参与者介绍研究目的和设计原则；然后，参与者佩戴 HMD 设备进行虚拟体验，分别采用四种原型设计方案进行情绪标注；在完成体验后，六位参与者针对标注设备的控制、标注信息的视觉反馈和标注任务的

认知负荷三个方面进行讨论并共同开发。本环节时长 45 分钟左右，全程录音，后期进行转录并采用开放编码（Open Coding）方法[144] 进行数据分析。主要讨论内容包括如下四个方面：

（1）标注信息的元素形式。六位研究员均认为边框或是渐变边框的方式不合适，P_3 表示“... 边框过于明显，会造成很大的干扰...”；P_5 指出“... 需要花费很多精力去关注视口周围显示的一圈颜色...”。同样，文本方式也得到了所有专家的否定，P_4 表示“... 在视口的中心放置一个标签严重遮挡了视野...”；P_1 提出可以将文字标签改为半透明形式叠加在视口区域。针对光束方案，由于光束的颜色强度会衰减，特别是在背景视频的光照或是主色调类似的情况下，颜色不易分辨，P_5 提出“... 如果光束的颜色更深一点会更好...”。在讨论的最后，参与者更倾向于采用无文字的配色方案，更加直观且较少地分散注意力。

（2）标注信息的元素位置。参与者普遍反对视觉信息占据整个视口或是视口的中心位置，相反更倾向于在特定区域呈现反馈信息。P_2 建议“... 在视口的四个边角区域提供反馈信息，会降低对注意力的分散...”。由于同时使用四个边角占据的区域较多，也没有增加额外信息，P_3 提出“... 可以根据标注的情绪类型所属的象限照亮相应的角落...”，但在这一方式中用户需要关注不同的视觉区域，会造成更多的注意力分散。经讨论，六位参与者明确将视觉反馈信息固定在视口的右下角区域。

（3）情绪强度信息。标注信息的另一关键要素是情绪强度的标注反馈，本环节提出两种方案：元素颜色的不透明度（用户所标注的情绪强度越大，元素的不透明度越高）；元素的尺寸大小（用户标注的情绪强度越大，元素的尺寸越大）。P_5 提出“... 大小的变化比颜色的透明度变化更直接...”，考虑到两种方案的特性差异，经讨论后仍然保留两版方案。

（4）即时查询功能。在配色方案中，元素颜色表示用户的情绪类型，用户在标注过程中可能会遇到忘记何种颜色表示何种情绪、何种情绪对应何种象限等问题，因此需要提供即时查询的功能解决这些问题。P_1 指出可以增加按钮控制事件为用户提供文字或是图片说明“... 设置一个开关，例如按下按钮显示...”；P_3 提出“... 可以把带有颜色信息的 *Circumplex* 模型面板叠加在虚拟场景中...”。基于此，本环节开发了即时查询功能，用户按住 Joy-Con 控制器上的指定按钮后，会在当前视口的中心区域叠加一个具有帮助信息的界面，包含 Circumplex“唤醒-效价”二维情绪模型图、四个象限的颜色信息和最具代表性的情绪关键词，如图3.4(b)所示。

(a) HaloLight（上）与 DotSize（下）方案　　(b) 即时查询功能界面

图 3.4　基于 HaloLight 与 DotSize 的标注信息可视化方案

综合考虑，本环节开发了两种标注信息可视化方案，HaloLight 和 DotSize。两种方案中可视化元素均固定在视口的右下角区域，元素的颜色表示所标注的情绪类型。在 HaloLight 方案中，元素形式为一个实心光晕弧，元素颜色的不透明度对应标注的情绪强度；DotSize 方案中，元素形式为一个实心圆，实心圆的尺寸大小对应标注的情绪强度。HaloLight 和 DotSize 方案如图3.4(a)所示。

3.3.3　评估体系

虚拟交互环境中实时连续的情绪标注任务，一方面需要确保能够获取准确有效的标注数据；另一方面，考虑到用户的标注对象是基于虚拟环境中主要任务所诱发的情绪状态，需要确保情绪诱发环境的可用性、以及情绪标注没有干扰主要任务中的用户体验。因此，本研究中标注方法的可用性从虚拟环境中的用户体验和标注数据的有效性两个方面进行评估，如图3.5所示。

对于虚拟环境中用户体验评估而言，晕动症[145] 和临场感[146] 是两个最基本的测量要素。VR 技术中最突出的问题之一是用户可能产生眩晕、恶心、呕吐等不适症状，即晕动症（Motion Sickness）[147]，研究采用标准化的 SSQ 量表（Simulator Sickness Questionnaire，SSQ）[148] 测量用户的晕动症；该量表包含 16 个小症状，评分 $i = 0, 1, 2, 3$，分别对应无症状、轻微、中等和严重。临场感（Presence）是指让用户产生“在场景中”的感觉，这也是 VR 环境中用户情绪反应的基础[149]。Schubert 等人编制的 IPQ 量表（Igroup Presence Questionnaire, IPQ）[150] 用于评估用户的临场感，包含空间沉浸感、卷入程度和真实感三个维度；IPQ 量表共有 13 个小项，采用 Likert-7 级评分（1=“完全不符合”；7=“完全符合”）。为了评估实时连续标注任务是否增加了用户虚拟

体验的认知负荷[137]，采用显式和隐式两类测量方法。显式测量法是指使用文件调查法和访谈法等方法获取用户主观反馈的测量方法，NASA-TLX 工作任务量表（NASA Task Load Index）[151] 广泛用于测量用户的认知负荷，量表包含心理需求、体力需求、时间需求、作业绩效、努力程度和挫折水平六个小项；访谈法可以根据预先设定好的会谈、提问结构直接获取用户的主观感受及体验。隐式测量法是指使用生理心理学信号对用户反馈进行客观测量的方法，瞳孔直径[152]、脑电信号[35] 及一些外周生理信号（如 EDA、HRV）[17] 等均可以反应认知负荷。

对于用户标注数据的有效性评估而言，主要采用假设检验等统计学方法。首先通过计算实时连续情绪标注数据与诱发素材情绪标签、以及体验后标注结果之间的相关性，判断标注数据的有效性；另一方面，检验同一用户针对不同情绪类型诱发素材标注结果之间的一致性、不同用户针对同一诱发素材标注结果之间的一致性，判断标注数据的可信度。

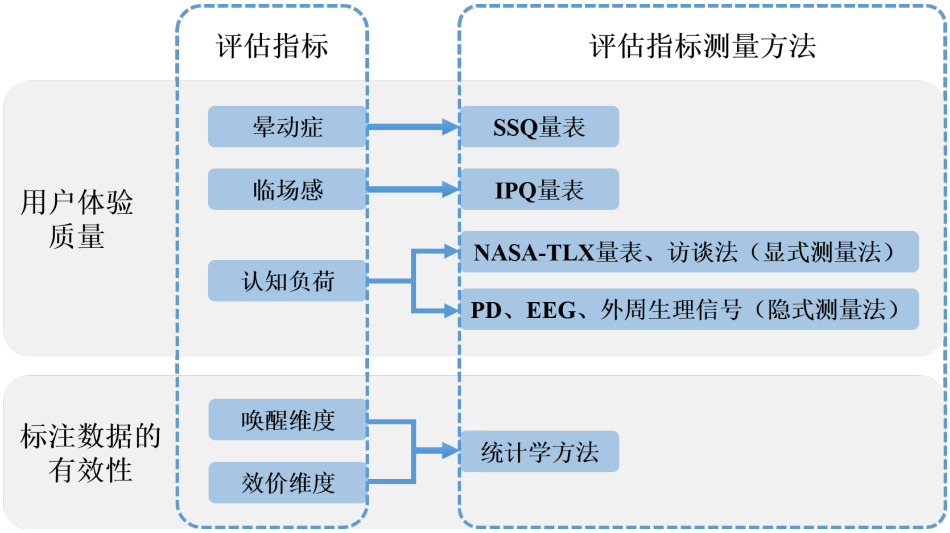


图 3.5　实时连续情绪标注方法评估框架

3.4　实时连续情绪测量实验

3.4.1　实验范式

虚拟交互环境中的实时连续情绪测量实验在可控的室内环境中进行，采用 2（两种标注方法：HaloLight 与 DotSize）× 4（四种情绪类型的视频：HAHV、LAHV、LALV、HALV）的被试内设计（Within-Subjects Design）。每种情绪类型选择两个视频，一共是八个视频（2 × HAHV、2 × LAHV、2 × LALV、2 × HALV）。用户佩戴 HMD 设备

观看全景视频的同时，采用 HaloLight 或 DotSize 方法实时连续标注情绪状态。本实验符合学校伦理委员会的相关要求，所有数据收集均征求了委员会和所有用户的同意。

3.4.1.1 诱发素材选择

本实验选取八个带有唤醒和效价评分的全景视频作为情绪诱发素材，视频均来自 Li 等人[153] 提出的数据集。Li 等人招募 95 位被试观看了 73 个全景视频，被试观看后通过 SAM 情绪评级给出唤醒与效价评分，该研究报告了目前唯一一个带有情绪标签的全景视频数据集。本实验从中选取了八个全景视频（每种情绪类型两个），见表3.2。视频分辨率为 4K（3840 × 1920 像素），等距圆柱体投影形式，均带有音频内容。由于数据集中的视频长度各不相同且均长于两分钟，被试在虚拟环境中长时间观看会产生晕动症和疲劳感[153,154]。为了避免这些问题，本研究从每个选定的视频中提取连续的 60 秒片段[76,155] 作为情绪诱发素材，并通过预实验验证选定的 60 秒内容能够诱发同样的情绪状态。

在预实验中，12 名研究员在虚拟环境中观看了八个视频的 60 秒片段并进行 SAM 情绪评级。针对 HAHV 类型的视频，第一轮选择的《Walk the Tight Rope》视频，效价维度评分偏低（$A = 4.17$）；第二轮分别测试了《Puppies host SourceFed for a day》与《Through Mowgli’s Eyes》，结合视频内容与评分结果（数据见表3.2），选择前者作为 HAHV 类型的诱发素材。组内相关系数（Intra-class Correlation，ICC）[156] 用于衡量情绪评分（$N = 12$）的观察者间信度（Inter-rater Reliability，IRR）。八个视频效价评分的 ICC 均值为 $ICC = 0.972, p < 0.05$，唤醒维度评分的 ICC 均值为 $ICC = 0.976, p < 0.05$，结果表明裁剪后的视频片段能够诱发同样类型的情绪，且八个视频片段的唤醒和效价评分在被试间具有很好的一致性[157]。

此外，表3.2中还计算了八个选定视频片段的空间感知信息（Spatial Perceptual Information，SpI）和时间感知信息（Temporal Perceptual Information，TpI）[158]。SpI 描述视频序列的空间信息量；TpI 表征视频序列的时间变化量。通过对选定视频序列在原始数据集中的唤醒-效价标签与 SpI、TpI 值的双向一致性 ICC 分析，发现二者之间没有相关性，表明本实验中诱发素材的时空特征对情绪标签的影响非常小。表3.2中也列出了选定视频的一些高级语义属性，诸如：室内/室外、视频类别、兴趣点；音频类别包括背景音乐、环境音、对话和画外音。

表 3.2　虚拟交互空间实时连续情绪测量实验采用的全景视频及属性信息

视频编号	情绪类型	数据集编号（V，A）	预实验（V，A）	视频名称	Youtube 链接	起始偏移量	SpI	TpI	音频类别	视频属性	内容描述
V0	Training	63 (6.36, 5.93)	/	NASA - Encapsulation & Launch of OSIRIS Rex	D7-AmamuJEA	7s	51.91	0.93	voice-over, bgm	indoor, docu-mentary	Documentary film on planning and execution of rocket launches
V1	HVHA	50 (7.47, 5.35)	(7.08, 6.08)	Puppies host SourceFed for a day	c7sA3EdXSUQ	0s	61.41	9.14	bgm	indoor, action, dogs	Viewers get up close with some puppies
V5	HVHA	52 (6.75, 7.42)	(6.83, 7.42)	Speed Flying	g6w6xkQeSHg	0s	65.04	12.88	dialog, bgm	outdoor, sport, pilot	Viewer follows a speed wing pilot as he glides past mountain
V3	LVHA	21 (3.20, 5.60)	(2.58, 6.83)	Zombie Apocalypse Horror	pHX3U4B6BCk	65s	55.98	2.61	dialog, ambience, bgm	indoor, film, zombies	Film following some soldiers defending against zombie attack
V7	LVHA	68 (4.40, 6.70)	(4.42, 7.17)	Jailbreak 360	vNLDRSdAj1U	127s	46.78	2.25	dialog, ambience, bgm	indoor, action, criminal	Short film depicting a jailbreak from closed-circuit cameras
V2	HVLA	38 (6.13, 1.80)	(8.08, 1.91)	Mountain Stillness	aePXpV8Z10Y	10s	39.42	0.97	bgm	outdoor, tour, mountain	Atmospheric shots of Canadian snowy mountains
V6	HVLA	32 (6.57, 1.57)	(7.67, 1.50)	Malaekahana Sunrise	-bIrUYM-GjU	0s	47.34	0.36	ambience	outdoor, tour, sunrise	Viewer sees the sun rising over the horizon at a beach
V4	LVLA	14 (2.53, 3.82)	(2.42, 4.17)	War Zone	Nxxb_7wzvJI	3s	62.99	1.54	voice-over, ambience, bgm	outdoor, film, people	Journalistic clip of a war torn city
V8	LVLA	19 (2.73, 3.80)	(2.17, 3.17)	The Nepal Earthquake Aftermath	5tasUGQ1898	41s	76.11	2.07	voice-over, ambience, bgm	outdoor, film, buildings	Short film on the effects of an earthquake in Nepal
abandon	HVHA	69 (6.46, 6.91)	(4.17, 7.00)	Walk the tight rope	JtAzMFcUQ90	10s	/	/	/	/	Viewer experiences walking a tight rope over a canyon
abandon	HVHA	73 (6.27, 6.18)	(5.50, 6.58)	Through Mowgli’s Eyes	bUiP-iGN6oI	13s	/	/	/	/	Short film with a conversation between an ape and a boy

3.4.1.2 被试招募

本实验共招募了 32 位被试，男女比例为 1:1，年龄区间在 18-33 岁之间（M=25，SD=4.0）。所有被试均来自附近的大学，多种国籍，人口统计学信息见表3.3。其中，62.5% 的被试有至少一次的 VR 体验，所有被试均了解全景视频，且没有视觉、听觉或是运动障碍。

表 3.3 实验被试的人口统计学信息

	年龄	性别	VR 体验次数	职业
人数	18-21：6		0 次：12	
	22-25：13	女性：16	5 次：15	教授：5
	26-29：9	男性：16	5-20 次：3	学生：27
	30-33：4		20 次：2	

3.4.1.3 系统架构

本实验的系统架构如图3.6所示，每个部分的详细介绍如下：

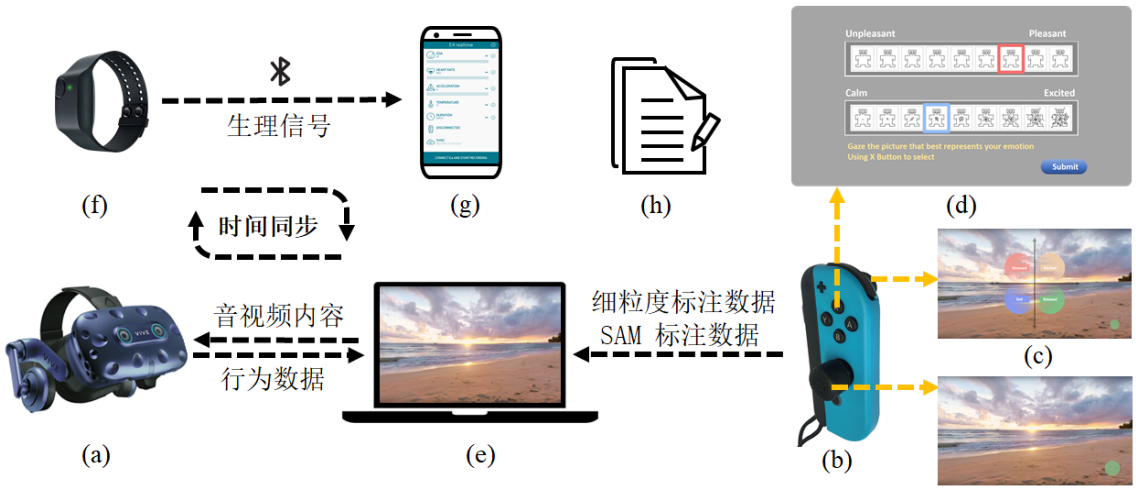


图 3.6 实时连续情绪测量实验系统架构图

（1）本实验选择 HTC VIVE Pro Eye HMD 作为头显设备，如图3.6(a) 所示。该设备单个目镜的分辨率为 1440 × 1600，双目分辨率为 2880 × 1600，视场角为 110 度，刷新率 90Hz。HMD 内置有 Tobii Pro 眼动追踪仪，双目凝视数据的输出频率为 120Hz，精度为 0.5 度。实验中，被试佩戴该设备观看选定的全景视频，并能够自由转动头部

选择观看方向，声音内容通过 HMD 的耳机传输。设备同步记录被试的头部运动数据和眼睛凝视数据，频率为 120Hz（见表3.4）。

（2）本实验采用带有摇杆的无线数字游戏控制器 Joy-Con 作为情绪标注设备，如图3.6(b) 所示。为了提升操作灵活度，在摇杆头处增加一个 11 毫米高的帽子，以延长摇杆长度。摇杆头的运动映射在“唤醒-效价”二维情绪空间，其中 X 轴对应情绪效价值，Y 轴对应情绪的唤醒值。摇杆头与摇杆中心的距离越远，表示情绪的强度越大。根据人体运动控制学研究[78,79]，情绪标注数据的采样频率为 10Hz（见表3.4）。

（3）图3.6(c) 所示为“即时查询”功能。在标注过程中，被试如果忘记何种颜色对应何种情绪类型，可以通过“即时查询”功能快速查看。按住 Joy-Con 控制器上的“Trigger”按钮后，会在当前视口的中心区域显示帮助界面，包含了最具代表性的情绪关键词，如图3.4(b)所示。

（4）图3.6(d) 所示为虚拟环境中内嵌的 SAM 情绪标注面板（Within-VR SAM Rating）。被试观看完每段视频后，需要给出 SAM 情绪评分。由于情绪是在虚拟环境中产生，因此情绪评分也应该在 VR 中进行[117]。该面板带有 Likert-9 级评分的唤醒和效价刻度值，唤醒维度值从“冷静”（1）到“兴奋”（9），效价维度值从“不开心”（1）到“开心”（9）。被试通过注视其中一个表示刻度值的图片进行选择，使用 Joy-Con 控制器上的“X”按钮完成自我报告。

（5）本实验的开发环境为 Unity 引擎[2]（版本号 2018.4.1f1），在引擎中构建了一个自定义场景来播放全景视频和音频内容，并采用 HaloLight 和 DotSize 方法实时反馈标注结果。全景视频以等距圆柱体形式投影在天空盒上，相机固定在球体中心。本实验采用 Tobii Pro SDK[3]从 HMD 中收集数据、SteamVR SDK[4]提供 VR 支持。项目运行环境为 2.2 GHz Intel i7 外星人系列笔记本电脑，显卡为 Nvidia RTX 2070。

（6）本实验采用 Empatica E4 手环[5]收集被试的外周生理信号，如图3.6(f) 所示。E4 手环是一个轻量级、简单易用的无线测量设备，能够有效地收集用于情绪识别的高质量信号[159]。该设备能够收集 SKT 数据和 EDA 数据；三轴加速度传感器能够获取手臂运动数据（Acceleration，ACC）；光电容积描记（Photoplethysmograph，PPG）传感器能够获取血容量脉冲（Blood Volume Pulse，BVP）数据，手环内嵌的算法可以从 BVP 中计算心率（Heart Rate，HR）和心脏各次跳动之间的时间间隔（Inter-beat Internal，

[2]https://unity.com/

[3]http://developer.tobiipro.com/unity/unity-getting-started.html

[4]https://store.steampowered.com/app/250820/SteamVR/

[5]https://www.empatica.com/en-int/research/e4/

表 3.4　实验设备采集的用户数据类型及采样频率

设备	采集的数据（简称）	采样频率（Hz）
HTC VIVE Pro Eye HMD	头部运动数据（HM）	120
	眼部运动数据（EM）	120
	瞳孔直径（PD）	120
Joy-con 控制器	效价度情绪标注（V）	10
	唤醒度情绪标注（A）	10
Empatica E4 手环	三轴加速度（ACC）	32
	血容量脉冲（BVP）	64
	皮肤电活动（EDA）	4
	心率（HR）	1
	心脏各次跳动之间的时间间隔（IBI）	\
	皮肤表面温度（SKT）	4

IBI）。各项生理指标的采样频率见表3.4。为了降低手臂运动伪影的干扰，被试在非惯用手上佩戴手环设备。

（7）本实验采用移动设备（(Nexus 5, 32GB, 5 I nches, 1920 × 1080）从 E4 手环中获取被试的生理信号相关数据，数据通过蓝牙传输，手环的时间戳通过 NTP 服务器[6]与实验用的电脑同步。

（8）本实验采用标准验证问卷对 VR 环境中被试的晕动症、临场感和任务负荷进行主观测量，分别采用 SSQ 量表、IPQ 量表和修改后的 NASA-TLX 工作任务量表[151]。

3.4.2　实验流程

实验流程如图3.7所示，整个流程持续时长约 50 分钟，具体流程介绍如下：

（1）在实验开始之前，被试需要签订知情同意书（Consent form）并填写有关背景信息的调查问卷（Background）；实验员向被试介绍实验内容和实验步骤，并解释情绪模型中的“唤醒”与“效价”维度含义以及 HaloLight 与 DotSize 情绪标注方法。在所有实验相关问题解决之后，被试填写实验前的 SSQ 问卷（SSQ 0）。

（2）在设备校准环节，实验员帮助被试测量瞳孔间距（Inter-Pupillary Distance，

[6]android.pool.ntp.org/

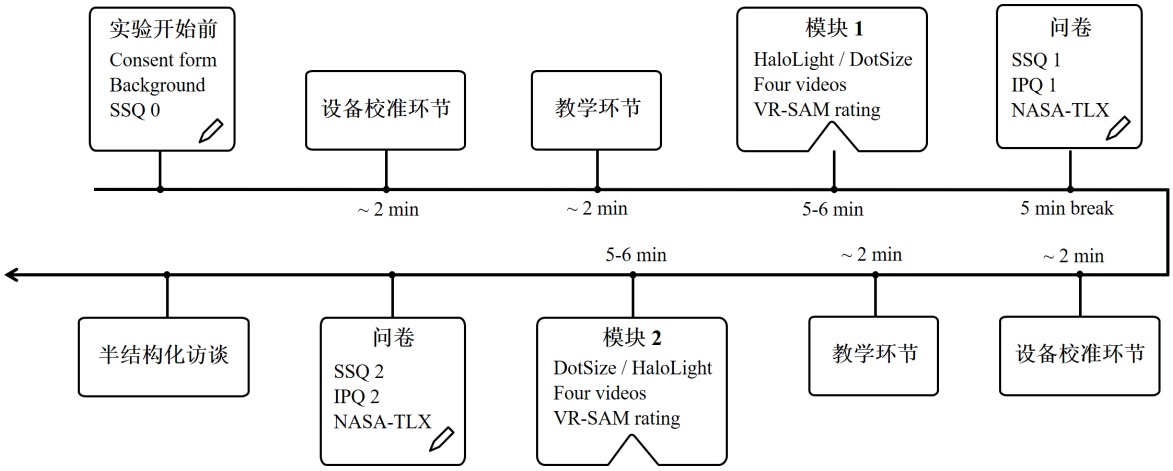


图 3.7　实时连续情绪测量实验流程

IPD），用于设置 HMD 两个目镜的距离。被试在非惯用手上佩戴 E4 手环，完成信号校准和采样测试。被试坐在转椅上并佩戴头显设备，调节座椅和 HMD 至舒适位置。根据 VIVE Pro Eye 的设置说明[7]进行眼动设备校准。被试每次摘掉后重新佩戴实验设备时，都需要进行该环节校准，即在模块 1 和模块 2 之前。

（3）在教学环节，被试佩戴 HMD 观看一个纪录片类型的全景视频（表3.2中的 V0），实验员以口述形式指导被试通过座椅旋转与头部运动和虚拟环境中的内容进行交互，帮助被试熟练使用 HaloLight 或 DotSize 标注方法。该环节发生在模块 1 和模块 2 之前。

（4）主实验环节包含两个模块，在每个模块中，被试观看四段不同情绪类型的全景视频，并采用 HaloLight 或是 DotSize 方法实时标注情绪状态。为了平衡 HaloLight 和 DotSize 的顺序影响，16 位被试在模块 1 中采用 HaloLight 方法进行情绪标注，模块 2 中采用 DotSize 进行标注；其余 16 位被试在模块 1 采用 DotSize 进行情绪标注，模块 2 采用 HaloLight 进行标注。实验采用部分析因设计（Fractional Factorial Design）[160]平衡两个模块中每种情绪类型使用的视频和不同情绪类型的视频播放顺序的影响。此外，为了消除被试在观看视频时因不同的初始观看位置对情绪造成的影响，在视频播放前 HMD 中会呈现带有一个白色方块的黑色场景，被试找到方块并凝视五秒后，方块消失同时开始播放视频。在预实验阶段，曾尝试通过固定转椅初始位置等方法统一用户的起始点，但小方块方法能够统一用户的眼部注视点。实验员在实验开始之前也向被试介绍了这一步骤。

（5）在主实验环节，同步收集被试通过摇杆设备连续报告的情绪状态、HMD 实

[7]https://www.vive.com/us/support/vive-pro-eye/category_howto/calibrating-eye-tracking.html

时捕获的被试头部运动与眼部运动数据、E4 手环记录的被试生理信号数据。为了避免不同情绪之间的叠加并降低被试观看全景视频的疲倦感，在两个模块之间设置了五分钟的休息间隔[63,161]。

（6）每观看完一个视频片段后，被试需要采用 VR 内嵌的 SAM 面板提交唤醒与效价两个维度的情绪评分。在每个模块完成之后，被试在实验员的帮助下摘掉 HMD 设备并填写 SSQ、IPQ、NASA-TLX 问卷。在两个模块均完成之后，被试参与一个半结构化访谈，访谈包括五个关于用户体验的问题。

3.5 实验结果与讨论

3.5.1 标注数据结果

本实验中，实验系统以 10Hz 频率记录摇杆头位置的 X 轴数值 u 和 Y 轴数值 v，范围是 $[-1,1]$，且摇杆头的运动在一个圆形区域中。唤醒与效价维度的 SAM 标注数据范围均为 $[1,9]$。因此需要对实时连续情绪标注数据进行预处理，使其与 SAM 标注数据结构一致。首先，采用简单拉伸方法（Simple Stretch Method）[162] 将标注数据 u 与 v 映射至方形区域中，计算如下：

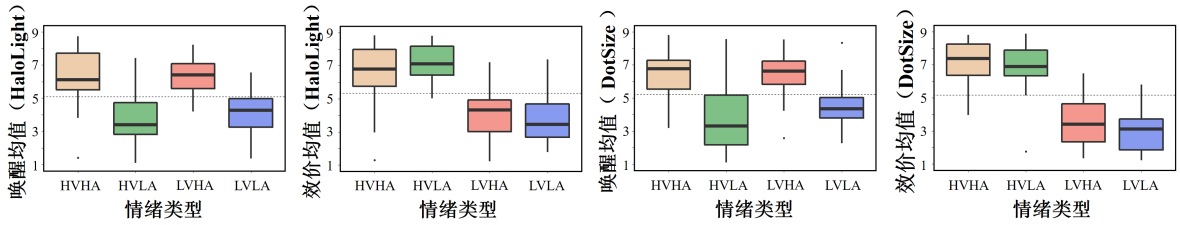
$$x=\begin{cases}0 & u=0\\ sgn(u)\sqrt{u^2+v^2} & |u|\geq|v|\\ sgn(v)\frac{u}{v}\sqrt{u^2+v^2} & |u|<|v|\end{cases}\qquad y=\begin{cases}v & u=0\\ sgn(u)\frac{v}{u}\sqrt{u^2+v^2} & |u|\geq|v|\\ sgn(v)\sqrt{u^2+v^2} & |u|<|v|\end{cases}\tag{3.1}$$

其中，$sgn(u)$ 用于提取 u 的符号。然后，将 x 和 y 的值由 $[-1,1]$ 缩放至 $[1,9]$，计算如下：

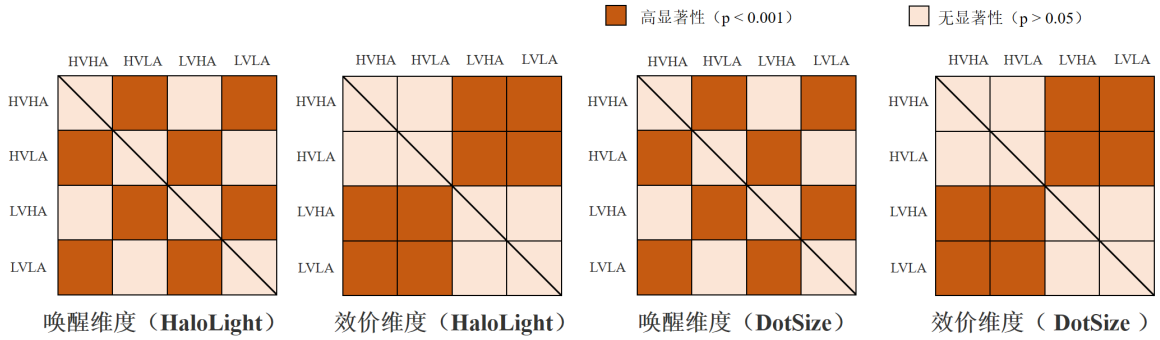
$$Valence=4*x+5\qquad Arousal=4*y+5\tag{3.2}$$

计算所有被试在 HaloLight 和 DotSize 两种方法中唤醒与效价维度的情绪标注均值，图3.8(a)所示为 32 个被试观看四种类型视频的情绪标注的效价-唤醒（V-A）均值分布。采用 Shapiro-Wilk 检验（W 检验）分析数据的正态性。W 检验结果表明 HaloLight 与 DotSize 的 V-A 标注均值均不符合正态分布（$p<0.05$），因此选择 Friedman 秩和检验方法进一步分析不同类型视频之间的情绪标注差异。HaloLight 方法在 V-A 情绪维度上的 Friedman 秩和检验结果为：效价（$\chi^2(3)=57.94, p<0.001$），唤醒（$\chi^2(3)=56.96, p<0.001$）；DotSize 方法的 Friedman 秩和检验结果为：效价

（$\chi^2(3)=71.44, p<0.001$），唤醒（$\chi^2(3)=43.39, p<0.001$）。上述结果表明视频的情绪类型对于 V-A 评分具有显著影响。为了明确任意两种类型的视频 V-A 情绪标注之间是否存在差异，采用 Bonferroni 成对比较方法进行事后检验[78,79]，结果如图3.8(b)所示，HaloLight 和 DotSize 在唤醒维度上，唤醒维度类型相反的视频之间具有显著性差异（$p<0.001$），唤醒维度类型相同的视频之间没有显著性差异（$p>0.05$）；HaloLight 和 DotSize 在效价维度上，效价维度类型相反的视频之间具有显著性差异（$p<0.001$），效价维度类型相同的视频之间没有显著性差异（$p>0.05$）。具有显著性差异的两类视频的效应值（Effect Size）范围是 $[0.600, 0.824]$。



(a) 四种情绪类型视频观看中唤醒与效价实时连续情绪标注均值箱线图



(b) 四种情绪类型视频观看中唤醒与效价实时连续情绪标注两两成对比较对称矩阵图

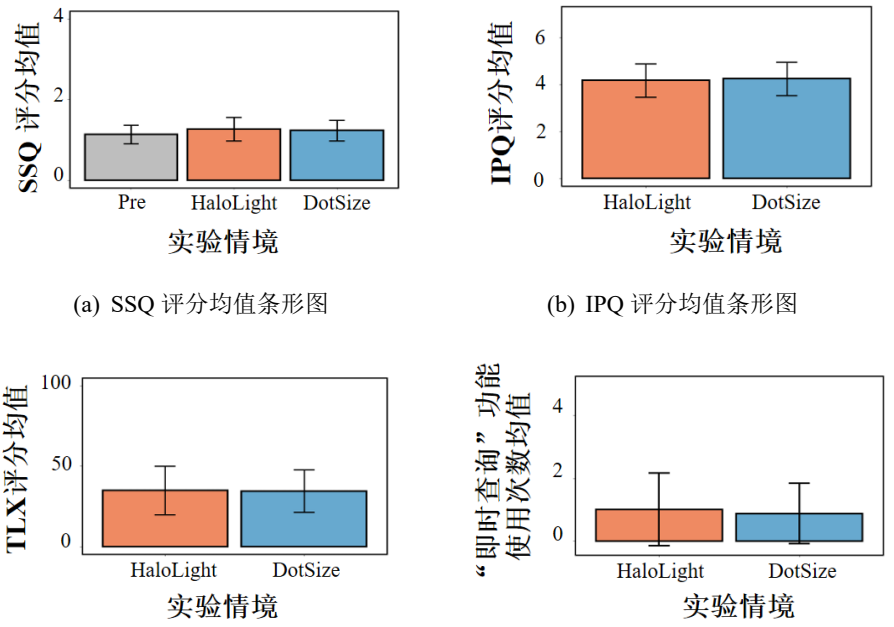
图 3.8　四种类型诱发素材的实时连续情绪标注均值结果

为了评估不同标注方法标注结果之间的一致性，采用双向混合、绝对一致且平均度量模型的 ICC 方法。ICC 均值结果表明 HaloLight 与 DotSize 两种方法的标注结果均值在效价维度上具有很好的一致性（$ICC=0.792, p<0.05$），在唤醒维度上具有较好的一致性（$ICC=0.606, p<0.05$）；HaloLight 与 SAM 方法在效价维度上具有很好的一致性（$ICC=0.855, p<0.05$），在唤醒维度上具有较好的一致性（$ICC=0.731, p<0.05$）；DotSize 与 SAM 在效价维度上具有很好的一致性（$ICC=0.909, p<0.05$），在唤醒维度上具有较好的一致性（$ICC=0.706, p<0.05$）。

3.5.2 用户体验结果

3.5.2.1 显式测量法结果

本小节分析了 HaloLight 和 DotSize 两种标注方法中 32 位被试的主观问卷量表（SSQ、IPQ、NASA-TLX）结果、“即时查询”功能使用次数与半结构化访谈结果。



(a) SSQ 评分均值条形图　(b) IPQ 评分均值条形图

(c) TLX 评分均值条形图　(d) “即时查询”功能使用次数均值条形图

图 3.9　用户体验主观报告数据结果

W 检验表明所有被试的 SSQ 评分均值不符合正态分布（$p < 0.001$），采用 Friedman 秩和检验比较实验前、HaloLight、DotSize 三个配对组间的差异性，结果表明实验前（$M = 1.139, SD = 0.229$）、HaloLight（$M = 1.268, SD = 0.290$）和 DotiSize（$M = 1.234, SD = 0.257$）三种情境下的 SSQ 评分没有显著性差异（$\chi^2(2) = 0.777), p = 0.106$）。W 检验表明 IPQ 评分均值符合正态分布（$p > 0.05$），执行配对样本 T 检验比较 HaloLight 和 DotSize 两种方法的差异性，结果表明 IPQ 评分在 HaloLight（$M = 4.181, SD = 0.710$）与 DotSize（$M = 4.250, SD = 0.711$）两种方法中没有显著性差异（$t(31) = 0.397, p = 0.694$）。W 检验表明“即时查询”功能使用次数不符合正态分布（$p < 0.05$），因此选择 Wilcoxon 符号秩检验比较 HaloLight 和 DotSize 两种方法的差异性，结果表明 HaloLight（$M = 1.008, SD = 1.153$）与 DotSize（$M = 0.875, SD = 0.963$）中被试使用“即时查询”功能的次数没有显著性差异（$Z = 0.801, p = 0.429$）。W 检验

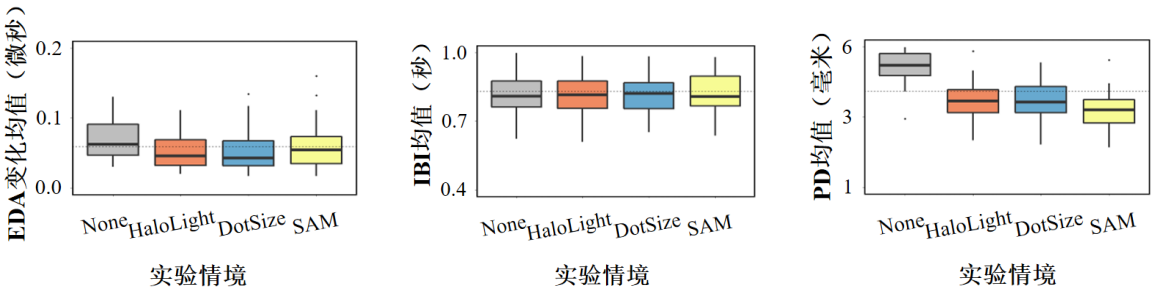
表明 NASA-TLX 任务量表的总分值符合正态分布（$p > 0.05$），配对样本 T 检验表明 HaloLight（$MD = 33.750, IQR = 20.417$）与 DotSize（$MD = 38.333, IQR = 19.791$）在任务负荷评分上没有显著性差异（$t(31) = 0.105, p = 0.917$）。因此，在 SSQ、IPQ、“即时查询”功能使用次数和 NASA-TLX 问卷方面，HaloLight 与 DotSize 两种方法的结果相似，如图3.9所示。

在实验后的半结构化访谈阶段，95% 的被试表示他们能够轻松处理好在虚拟环境中观看全景视频和实时连续情绪标注两项任务。其中一个问题是“*HaloLight 和 DotSize 两种标注方法，你更倾向于（喜欢）哪种？*”。13 位被试（43%）选择 HaloLight；15 位被试（47%）更倾向于 DotSize，其中的八位表示 HaloLight 占据了更多的视觉空间并影响到了观看体验。P_4 和 P_{14} 指出视频内容的偏好会影响对标注方法偏好的判断。P_4 说“*... 我用 HaloLight 标注的一个视频是滑雪主题，有严重的眩晕感，并且我不喜欢这项运动。但是在 DotSize 方法阶段，标注的一个视频是小狗主题，可爱的狗狗令我很开心，因此我也更倾向于 DotSize...*”此外，P_2 还提到她更喜欢 DotSize 是因为在第二模块更熟悉标注任务，因此由于顺序影响，DotSize 也带来了更好的印象。四位被试（12%）对于两种方法没有倾向，P_{32} 表示：“*... 如果是实心圆和透明度变化的组合，效果会更好...*”。

3.5.2.2 隐式测量法结果

本小节分析了无标注任务（None）、两个实时连续情绪标注模块（HaloLight/DotSize）、VR 内嵌的 SAM 评分（SAM）三种情境下被试外周生理信号（PD、EDA 变化、IBI）的差异性，结果如图3.10所示。在 None 情境下，被试无需执行任何操作，任务负荷会很低，并以此作为基线；在 SAM 情境下，被试需要在完成视频观看后给出 V-A 情绪标注；在 HaloLight/DotSize 情境下，被试在虚拟体验过程中实时报告情绪状态。因此实验假设在上述三种情境下被试的任务负荷从低到高依次是：None < SAM < HaloLight/DotSize。

瞳孔直径是从 HMD Tobii 眼动设备中直接获取的原始数据，单位是毫米，采样频率为 120Hz。四种情境下 PD 的均值与标准差分别为：$None = 4.777(0.687)$、$HaloLight = 3.473(0.572)$、$DotSize = 3.444(0.574)$、$SAM = 3.209(0.526)$。W 检验表明 PD 值不符合正态分布（$p < 0.05$），Friedman 秩和检验用于比较 PD 值在四个配对组间差异性，结果表明不同情境中的 PD 值具有显著差异性（$\chi^2(3) = 73.95, p < 0.001$）。采用



(a) EDA 变化均值箱线图　(b) IBI 均值箱线图　(c) PD 均值箱线图

图 3.10　HaloLight 和 DotSize 两种标注方法中用户生理信号及瞳孔直径结果

Bonferroni 成对比较方法进行事后检验，结果表明显著性差异存在于 None 与 HaloLight 之间（$Z = 5.841, p < 0.01, r = 0.730$）、None 与 DotSize 之间（$Z = 5.975, p < 0.01, r = 0.747$）、None 与 SAM 之间（$Z = 6.297, p < 0.01, r = 0.787$），而 HaloLight 与 DotSize 之间（$Z = 0.081, p > 0.05$）、HaloLight 与 SAM 之间（$Z = 1.947, p > 0.05$）、DotSize 与 SAM 之间（$Z = 1.846, p > 0.05$）没有显著性差异。Empatica E4 手环内置的 EDA/GSR 传感器能够获取 EDA 数据，采样频率为 4Hz。根据 Fleureau 等人的研究[163]，EDA 信号的一阶导数能够反应唤醒值的变化。采用截止频率为 2Hz 的三阶低通滤波器移除 EDA 数据伪影，将过滤后 EDA 信号的非负一阶导数作为 EDA 变化值[79]。四种情境下 EDA 变化的均值与标准差分别为：$None = 0.065(0.030)$、$HaloLight = 0.054(0.027)$、$DotSize = 0.053(0.029)$、$SAM = 0.060(0.034)$。W 检验表明 EDA 变化值不符合正态分布（$p < 0.05$），Friedman 秩和检验结果表明四种情境下 EDA 变化没有显著差异（$\chi^2(4) = 7.609, p = 0.055$）。IBI 数据测量被试两次心跳之间的时间间隔，单位是秒。Empatica E4 手环中的 PPG 可以收集 BVP 信号，采样频率为 64Hz，手环内置的处理算法消除了由于 BVP 信号噪声造成的不正确峰值[8]，并从中计算 IBI 序列。四种情境下 IBI 数据的均值和标准差分别是：$None = 0.825(0.097)$、$HaloLight = 0.838(0.099)$、$DotSize = 0.832(0.101)$、$SAM = 0.839(0.103)$。W 检验表明 IBI 值不符合正态分布（$p < 0.05$），Friedman 秩和检验表明四种情境下的 IBI 值没有显著性差异（$\chi^2(3) = 3.902, p = 0.272$）。

[8]https://support.empatica.com/hc/en-us/articles/360030058011-E4-data-IBI-expected-signal

3.5.3 标注方法有效性讨论

针对虚拟环境中用户体验评估的研究，Krüger 等人[125] 指出被试在虚拟环境中填写问卷量表比摘掉 HMD 填写问卷耗时短，且降低了对被试临场感的干扰。Schwind 等人[164] 和 Putze 等人[124] 的研究结果也表明 VR 内与 VR 外被试的临场感没有显著性差异，但方差的一致性具有显著性差异，即在虚拟环境中填写问卷能够减少对被试临场感的干扰且避免偏差。本实验中 SAM 情绪评级在虚拟环境中进行，而 SSQ、IPQ、NASA-TLX 问卷量表在摘掉 HMD 后立即完成。考虑到三个问卷小项数量多、题目阅读难度大，因此没有嵌入在虚拟环境中。

本实验中被试的 SSQ 评分均值相比于先前观看全景视频的研究[165] 非常低；并且实验前、HaloLight 和 DotSize 三种情境下的 SSQ 评分没有显著性差异。这表明 HaloLight 和 DotSize 两种实时连续的情绪标注方法没有给被试带来严重的晕动症。可能的原因有本实验选择的诱发素材时长短、视频拍摄的相机运动速度慢、被试采用坐在转椅上的观看模式等。IPQ 问卷分析结果表明 HaloLight 与 DotSize 两种方法中被试的临场感没有显著性差异，Subramanyam 等人[166] 的研究指出 3DoF 媒介具有较好的 IPQ 评分，本实验中的 IPQ 分数与此一致。在任务负荷方面，HaloLight、DotSize 与 SAM 标注方法的 NASA-TLX 量表和生理信号（PD、EDA 变化、IBI）测量结果均没有显著性差异。相比于 Zhang 等人[79] 在移动设备环境中的实时连续情绪标注研究，本实验中 NASA-TLX 分数更低。这表明相比于传统的事后情绪评分方法，HaloLight 和 DotSize 均没有增加被试的认知负荷。值得注意的是，被试的瞳孔直径在 HaloLight、DotSize 与 SAM 三种情境下均与 None 情境下具有显著性差异。Pfleging 等人[167] 与 Zhu 等人[168] 的研究指出，PD 值一方面受到用户情绪体验影响，另一方面也会受到用户所处环境的整体光强度影响，光线暗的环境中被试 PD 值会变高。在 None 情境下，HMD 中呈现纯黑色场景，因此被试的 PD 值比高于其他情境。此外，78% 的被试平均每个视频使用了一次“即时查询”功能，并且 HaloLight 与 DotSize 模块中的使用次数没有显著性差异。半结构化的访谈结果也表明两种方法在双重任务中都方便易用。

本实验中八段诱发素材的实时连续情绪标注结果与原始数据集中的 V-A 标签一致，同时也与预实验中的标注结果一致。统计学分析结果表明情绪类型相同的视频之间实时连续情绪标注结果均没有显著性差异（$p > 0.05$），情绪类型相反的视频之间标注结果均具有高度显著性差异（$p < 0.001$）。不同标注方法标注结果的一致性检验结果表明：HaloLight 与 DotSize 两种方法的 V-A 标注具有很高的一致性；HaloLight/DotSize

与 SAM 情绪标注结果高度一致；SAM 情绪标注与原始数据集中的 V-A 标签高度一致（V：$ICC = 0.982, p < 0.05$；A：$ICC = 0.941, p < 0.05$）。此外，Voigt-Antons 等人[126] 的研究在虚拟环境中嵌入二维情绪坐标网格图获取用户的情绪报告数据，本实验中 HaloLight、DotSize 与 SAM 的标注结果一致性远高于该研究报告的数值。

综上，（1）实时连续的情绪标注实验获得了较好的用户体验，HaloLight 和 DotSize 方法均没有增加用户的晕动症和认知负荷，也没有干扰用户的临场感；（2）HaloLight 与 DotSize 两种实时连续的情绪标注方法均能够在虚拟空间中收集用户准确有效的连续情绪标注数据；（3）HaloLight 与 DotSize 两种方法的标注结果没有显著性差异，能够将其整合用于进一步分析。

3.6 本章小结

本章探讨了实时连续情绪标注的研究背景和研究现状，从人机交互的设计准则出发，首先聚焦虚拟环境中实时连续情绪标注方法的三个设计原则：（1）考虑基于头戴显示器设备的虚拟交互环境；（2）考虑实时连续情绪标注设备的人体工程学性能；（3）考虑多重任务引起的注意力分散；基于此选择 HTC VIVE Pro Eye HMD 呈现高质量虚拟内容、轻量级的 Joy-Con 无线摇杆控制器作为输入设备，并通过以用户为中心的多领域专家共同设计方法形成 HaloLight 与 DotSize 两种标注信息可视化方案。研究还建立了实时连续情绪标注方法的评估体系，由用户体验质量和标注数据的有效性两个方面组成，其中用户体验质量采用显式和隐式两类方法分析晕动症、临场感和认知负荷三个指标，标注数据的有效性从唤醒和效价两个维度采用统计学方法进行评估。本章构建了虚拟交互环境中实时连续情绪诱发及测量实验范式，提出基于 HaloLight 和 DotSize 的情绪诱发实验场景和情绪数据采集系统。通过实时连续情绪测量实验，验证了 HaloLight 与 DotSize 两种标注方法的可用性。实验结果表明 HaloLight 与 DotSize 两种实时连续的情绪标注方法能够收集虚拟环境中用户精确有效的唤醒和效价维度情绪标签；两种方法在晕动症、临场感和任务负荷方面没有显著性差异，且没有增加用户的晕动症和认知负荷、也没有干扰用户的临场感。因此，本章构建的方法可用于获取虚拟体验中用户精确有效的实时连续情绪 Ground-Truth 标签。

第 4 章　多模态情绪数据集构建方法

4.1　引言

在情感计算领域，精确有效的情绪数据至关重要，但数据的收集往往是一个漫长、困难或昂贵的过程，简化这一过程的重要替代方法是使用公开有效的数据集。现有的多模态情绪数据集同时包含了用户的显式情绪自我报告与隐式情绪响应（生理信号、动作表情等），如 MAHNOB-HCI[77]、DEAP[76]、CASE[78] 等。这些数据集已用于开发情绪识别算法、训练模型自动检测情绪[17,169]，但用户的情绪诱发均是采用二维视频或音频片段等非沉浸式情境。

伴随着 VR 技术的快速发展和 HMD 设备迅速商业化，全景视频逐渐涌入日常生活[170,171]；相比于传统媒体，全景视频能够带来全沉浸式交互体验[170]。为了更好地理解用户在虚拟体验中的情绪状态，研究聚焦用户在体验过程中的生理及行为数据。常用于情绪认知测量的生理指标有 EEG、HRV、EDA 等，这些指标已用在 VR 游戏[172]、用户体验质量评估[173]、驾驶模拟体验[174] 和构建虚拟情感系统（AVRS）[175] 研究中。虚拟体验中一个重要特性是个体与情绪诱发场景的交互方式不尽相同，在这一层面，若干研究分析了用户头部运动与眼部运动和临场感[176]、焦虑感[177] 及唤醒与效价情绪维度[63,178] 之间的相关性。但是，绝大多数的研究均没有公开实验相关数据。

在虚拟环境中，现有的数据集主要关注用户视觉行为模式[52,179]、或是视觉行为评估[180]。对于情绪研究，Li 等人[63] 提出了虚拟环境中首个带有 SAM 情绪标签的 VR 数据集，但没有考虑情绪的时变性。上一章的研究提出了虚拟环境中实时连续情绪标注数据的重要性与关键性，本章构建了虚拟环境中首个带有实时连续情绪 Ground-Truth 标签的多模态公开数据集，包含用户体验过程中的视觉行为数据、外周生理信号、连续情绪标注及体验后报告，如图4.1所示。本章的研究内容分为如下三个方面：

（1）基于3.4小节的实时连续情绪测量实验，研究公开了一个虚拟环境中实时连续的生理和行为情绪标注数据集（CEAP-360VR），包含 32 位被试观看八个全景视频的诱发素材与体验后问卷数据（SSQ、IPQ、NASA-TLX），观看过程中实时采集的连续情绪标注数据、头部和眼部运动数据、瞳孔直径数据、外周生理信号（ACC、EDA、SKT、BVP、HR、IBI），以及数据获取、处理和验证脚本。数据集链接为 https://github.com/cwi-dis/CEAP-360VR-Dataset。

（2）采用统计学研究方法，验证 CEAP-360VR 数据集多模态用户数据的有效性并介绍了数据集在情绪识别中的多种使用方式。在对多模态用户数据预处理之后，基于所有被试的标注轨迹与推论统计分析研究不同诱发素材中的实时连续情绪标注；分析不同情绪类型的全景视频观看过程中瞳孔直径状态，以及外周生理信号分布情况。

（3）采用机器学习技术进行一系列基线分类实验，进一步验证 CEAP-360VR 数据集多模态用户数据的有效性和可信度。选择四种机器学习方法与两种深度学习方法，在唤醒-效价两个维度情绪标签执行二分类、三分类、五分类三项分类任务，以及消融实验，进行情绪的分类、识别和预测。

基线分类实验结果表明，随机森林分类器在 2s 时长片段下具有良好的分类准确率：对于用户依赖评估模型，二分类任务的效价和唤醒维度准确率分别是 68.45%、71.33%，三分类任务的效价和唤醒维度准确率分别是 60.42%、62.38%；对于用户独立模型，二分类任务的效价和唤醒维度准确率分别是 66.80%、64.26%，三分类任务的效价和唤醒维度准确率分别是 49.92%、52.20%。此外，消融实验结果表明仅使用行为数据模态或生理信号模态产生了合理的识别精确度，同时使用这两个模态能够提升分类性能。

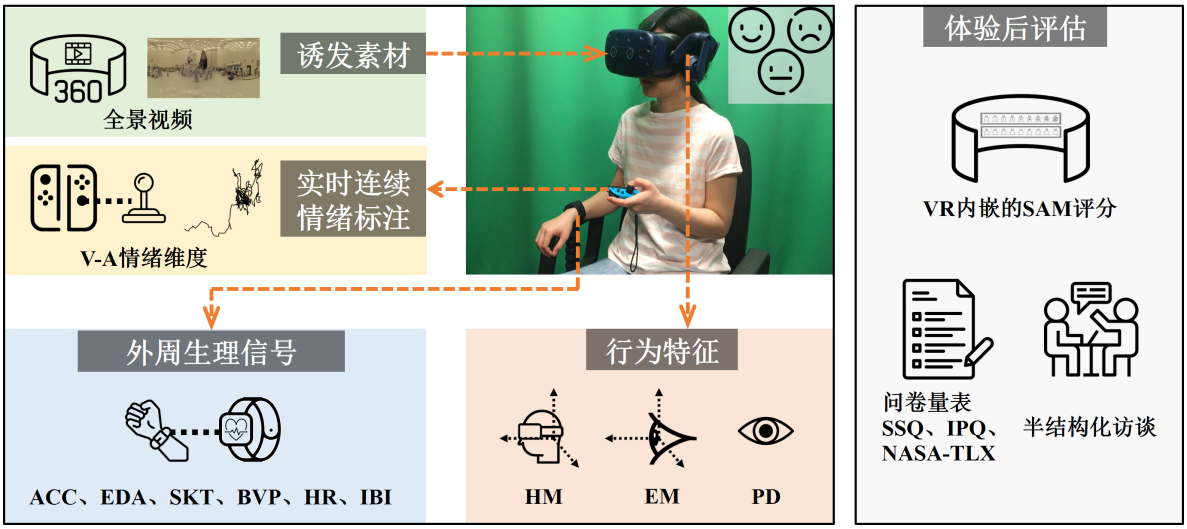


图 4.1　CEAP-360VR 多模态情绪数据集内容

4.2　相关工作

情感计算领域的很多研究构建了公开的多模态情绪数据集，最常见的应用场景为二维视频观看。Soleymani 等人[77] 收集了 27 位被试观看 34 个视频和图片时的生理

信号，包括 ECG、EDA、EEG、RESP、SKT；提出的 MAHNOB-HCI 数据集还包含被试的面部表情视频、眼部凝视数据和离散的唤醒-效价-支配、预测度与情绪关键词主观评价。Koelstra 等人[76] 采用隐式测量方法收集了 32 位被试观看 40 个视频片段时的生理信号，包括 EEG 和 BVP、EOG、ECG、EDA、SKT、RESP 六项外周生理信号；构建的 DEAP 数据集还包含唤醒-效价-支配维度、喜好与熟悉度主观评价数据。Miranda-Correa 等人[181] 提出的 AMIGOS 数据集包含 40 位被试观看 20 个长视频和短视频时的生理信号（EEG、ECG、GSR），以及情绪级别的内部评价（问卷量表）和外部评价（记录被试肢体动作的视频）。Subramanian 等人[182] 提出的 ASCERTAIN 数据集获取了 58 位被试生理信号（ECG、EDA 和 EEG）和面部动作数据，以及唤醒-效价、喜好度、关注度、熟悉度和“大五”人格的离散维度主观评级数据。最近，Sharma 等人[78] 收集了 30 位被试观看八个带有情绪标签的二维视频时的情绪响应，提出的 CASE 数据集包含已完成时间同步处理的生理信号（ECG、BVP、GSR、EMG、SKT、RESP）和唤醒-效价两个维度的实时连续情绪标注数据。这些数据集可用于构建精确的情绪标签、或开发自动情绪识别算法[169]，但情绪诱发场景均为非沉浸式情境。

MacQuarrie 等人[170] 和 Egan 等人[173] 对佩戴 HMD 设备、CAVE 环境和平面显示器三种模式下观看全景视频的用户体验质量从临场感、交互度、可用性和晕动症四个方面进行比对分析，结果表明相比于传统模式，HMD 设备可以带来更好的用户体验。但是，虚拟环境中公开可用的情绪数据集非常稀缺。Li 等人[63] 提出了虚拟环境中首个公开的情绪数据集，包含 93 名被试观看 73 个全景视频时的头部运动特征和体验后 SAM 情绪标注数据。Egan 等人[173] 首次在虚拟环境用户体验质量评估的研究中使用皮肤电和心率生理信号。Marín-Morales 等人[35] 采集了被试的 EEG 和 HRV 数据、及唤醒-效价维度情绪感知数据，结果表明虚拟环境能够诱发用户的情绪状态，且用户生理信号与情绪之间的关系和物理环境结果一致。但是，虚拟环境中现有的生理信号和情绪研究均基于研究者自己收集的数据，尚未形成公开的数据集，这使得其他研究员无法进行结果复现和比对分析[96]。

值得注意的是，在 CEAP-360VR 数据集公开之后，Tabbaa 等人[57] 提出了虚拟环境中用于情绪识别的多模态数据集 VREED，包含用户在观看全景视频后的自我报告问卷、预处理的眼部运动数据、ECG 与 GSR 生理数据，但该数据集仍没有考虑用户情绪状态的实时性和连续性。相比于此，本研究公开的 CEAP-360VR 数据集包含实时连续情绪标注数据，能够在细粒度层级上进行虚拟环境中情绪状态及生理和行为反馈

研究。

4.3 CEAP-360VR 数据集

基于实时连续情绪测量实验（见3.4小节），本研究公开了一个虚拟环境中实时连续的生理和行为情绪标注数据集（CEAP-360VR）。该数据集包含 32 位被试观看八个全景视频时多模态的原始数据和预处理后数据，为每个被试和视频分配一个唯一标识符，分别记为 $P1$–$P32$ 和 $V1$–$V8$，在数据集中采用字母 PXX 表示被试的编号（其中 XX 是集合 $\{1, 2, ...32\}$ 中的自然数），采用 VXX 表示视频的编号（其中 XX 是集合 $\{1, 2, ...8\}$ 中的自然数）。所有数据均采用 JSON（JavaScript Object Notation）格式存储，方便外部访问和处理[183]。同时，数据集还带有数据处理脚本和数据集基线验证脚本，以及相关方法描述文件。数据集的公开链接为 https://github.com/cwi-dis/CEAP-360VR-Dataset。

CEAP-360VR 数据集结构如图4.2所示，包含 $CEAP-360VR$ 数据文件夹、$CEAP-360VRDatasetDescription.pdf$ 数据集描述文件、数据集许可说明 $License.txt$ 文件与数据集的 $ReadMe.md$ 文件。CEAP-360VR 数据集遵循 Creative Commons At tribution-NonCommercial-NoDerivatives 4.0 International License（CC BY-NC-ND 4.0）许可文件[1]。数据集描述文件介绍了数据详细内容、数据获取和数据处理关键步骤。

4.3.1 诱发素材与问卷数据

$1\_Stimuli$ 文件夹由 $VideoThumbNails$ 子文件夹和 $VideoInfo.json$ 文件组成，其中 $VideoThumbNails$ 文件夹内是八个诱发素材具有代表性的缩略图（.jpg），可用作生成显著图时的底图等；$VideoInfo.json$ 文件包含了八个全景视频的详细信息：编号、名称、分辨率、时长、帧率、帧数、链接，以及原始数据集中唤醒-效价维度评分均值与裁剪后视频在预实验中的唤醒-效价评分均值。$2\_QuestionnaireData$ 文件夹内容为 32 位被试在实验过程中填写的纸质问卷数据以及在 VR 环境中针对每个视频的 SAM 评分数据，每位被试的相关数据存储在 $PXX\_QuestionnaireData.json$ 文件内。其中 SSQ, IPQ and NASA-TLX 问卷数据分别采用如下范围记录 $[1, 4]$, $[-3, 3]$, $[1, 20]$。VR 内嵌的 SAM 评分数据包含唤醒-效价两个维度评分结果、八个视频播放顺序、外周视觉反馈顺序（其中 1 表示模块 1，2 表示模块 2），同时还存储了每个视频播放的

[1]https://creativecommons.org/licenses/by-nc-sa/4.0/

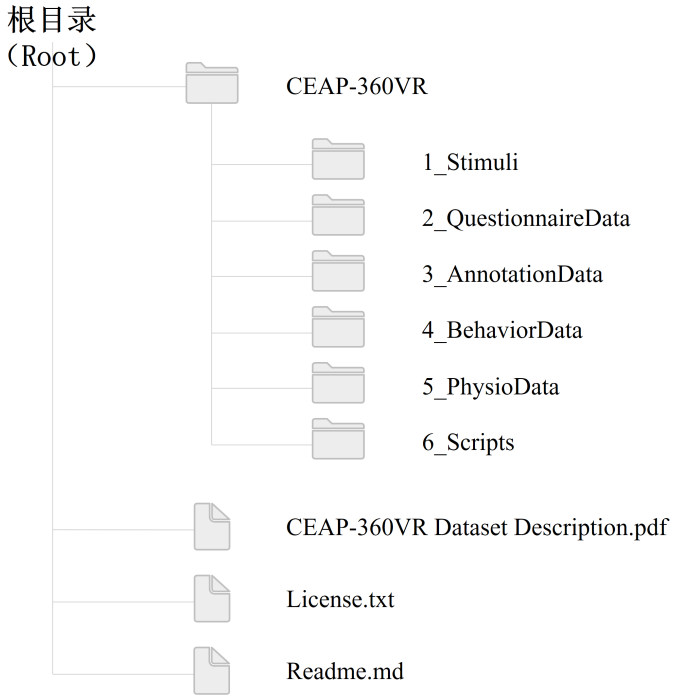


图 4.2　CEAP-360VR 数据集文件结构

起止时间戳，并通过以下 4 种形式记录在 json 文件中：本地 UTC/GMT 时间戳、秒级 Unix 时间戳，毫秒级 Unix 时间戳以及 HMD 设备时间戳，这些数据可用于对被试的行为数据和生理信号数据进行时间同步处理。

4.3.2　多模态用户数据

3_$AnnotationData$、4_$BehaviorData$、5_$PhysioData$ 三个文件夹的内容分别是 32 位被试观看八个视频中实时连续的情绪标注数据、头部与眼部运动及眼部瞳孔直径数据、外周生理信号数据。上述文件夹中的文件内容和对应的变量类型见表4.1，多模态用户数据预处理步骤如下：

（1）时间戳标定。基于 $PXX\_QuestionnaireData.json$ 文件中列出的视频播放起始和结束时间戳，依次过滤出每个视频播放期间被试的所有采样数据，并为每个采样数据添加新的时间戳：精度定为毫秒，单位是秒，视频播放起始时间戳计为 0。

（2）原始数据转换。多模态数据采集自不同的设备或传感器，数据之间存在属性差异，需要采用不同的转换方法。对于实时连续情绪标注数据，采用简单拉伸方法将 Joy-Con 摇杆设备中获取的标注数据映射至方形区域后，对唤醒-效价两个维度的情绪数据调整至 $[1, 9]$，与 SAM 评级的数据范围一致，详细转换方法见3.5.1小节；对

表 4.1　CEAP-360VR 数据集主要文件类型与变量类型

多模态数据	文件类型	变量类型
实时连续情绪标注数据 3_AnnotationData	PXX_Annotation_RawData.json	Timestamp, X_Value, Y_Value
	PXX_Annotation_TransformedData.json	TimeStamp, Valence, Arousal
	PXX_Annotation_FrameData.json	TimeStamp, Valence, Arousal
行为数据 4_BehaviorData	PXX_Behavior_RawData.json	HM, EM, LEM, REM, (X,Y,Z); LPD, RPD
	PXX_Behavior_TransformedData.json	HM, EM, LEM, REM, (Pitch, Yaw); LPD, RPD
	PXX_Behavior_FramedData.json	HM, EM, LEM, REM, LPD, RPD
	PXX_Behavior_GazeFixationData.json	StartFrame, EndFrame, Pitch, Yaw
	PXX_Behavior_HeadScanPathData.json	PointID, Pitch, Yaw
外周生理信号数据 5_PhysioData	PXX_Physio_RawData.json	TimeStamp, ACC, EDA, SKT, BVP, HR, IBI
	PXX_Physio_TransformedData.json	TimeStamp, ACC, EDA, SKT, BVP, HR, IBI
	PXX_Physio_FrameData.json	TimeStamp, ACC, EDA, SKT, BVP, HR, IBI

于头部与眼部运动数据，HMD Tobii 眼动设备提取的采样点数据包含头部运动欧拉角 $rotation(x,y,z)$（x,y,z 值的范围是 $[0,360]$）、世界坐标系内左眼右眼和双眼凝视方向标准化向量（x,y,z 值的范围是 $[-1,1]$）、左眼和右眼瞳孔直径（LPD、RPD，单位是毫米），将头部运动和眼部运动数据转换为观看方向的经度和纬度，同时从眼部运动中计算眼部的注视和扫视数据，从头部运动中计算头部扫描路径数据，详细转换方法见5.3.1小节；对于外周生理信号数据，进行滤波降噪处理后对每个信号进行标准化处理，详细转换方法见4.4.3小节。

（3）数据重采样。多模态数据采集自不同的设备或传感器，不同类别数据的采样频率不同（见表3.4）。为了对所有被试间不同类别数据进行对齐和数据同步，首先给每个视频的每一帧标定时间戳，将所有数据重采样至与视频帧率一致。如果原始数据的采样频率低于视频帧率，采用线性插值方法通过将原始数据的离散样本拟合为一条直线来确定新采样时间戳对应的数据；如果原始数据的采样频率高于视频帧率，选择小于重采样时间戳的最大采样点作为新采样数据。

（4）数据存储格式。上述步骤中，（1）用于从所有采集设备记录的用户数据中获取视频播放阶段的用户原始数据；（2）用于将原始数据转换为易于分析的通用格式，并对部分数据进行降噪滤波处理；（3）用于将转换后的用户数据重采样至帧数据。本步骤将用户的实时连续情绪标注数据、头部和眼部行为数据以及生理信号数据，分别

存储为以下三个类别：原始数据 Raw，转换后数据 $Transformed$ 与重采样后的帧数据 $Frame$，具体文件类型与名称见表4.1。

4.3.3 相关脚本文件

6_$Scripts$ 文件夹内容是数据集中多模态用户数据获取、处理、分析和验证的相关源代码（Python 脚本），包含 1_$UnityProject$、2_$DataProcessed$ 与 3_$CEAP-360VR_Baseline$ 三个子文件夹。其中 1_$UnityProject$ 文件夹是实时连续情绪测量实验的项目工程文件，主要功能有播放全景视频、可视化标注信息、输出 Joy-Con 控制器数据以及记录头部运动和眼部运动相关数据等，开发环境为 Unity 2018.4.1f1，采用 C# 语言。2_$DataProcessed$ 文件夹包含实时连续情绪标注、头部和眼部行为、外周生理信号三类原始数据的数据预处理脚本，以及各模态数据的分析验证和可视化脚本。3_$CEAP-360VR_Baseline$ 文件夹的内容是多模态数据特征提取脚本，以及基于用户独立和用户依赖模型的机器学习与深度学习基线实验脚本。

4.4 用户数据统计学验证

本节围绕 CEAP-360VR 数据集中的实时连续情绪标注数据、瞳孔直径数据和外周生理信号数据（EDA、SKT、HR 与 IBI）开展描述性统计学验证，并对实验结果进行讨论。

4.4.1 情绪标注数据结果与讨论

3.5.1小节验证了 HaloLight 与 DotSize 两种方法针对四种情绪类型视频的唤醒-效价情绪标注数据有效性。本节融合两种标注方法的标注结果，分析 32 位被试在观看八个全景视频时实时连续情绪标注结果的有效性。

首先计算每个视频每一帧内所有被试唤醒和效价标注均值，生成的八条标注轨迹如图4.3所示。从图中可以看出，八个视频唤醒和效价标注结果与诱发素材的情绪标签一致；同属一种情绪类型的两个视频标注轨迹位于同一象限，这表明 CEAP-360VR 数据集中用户连续标注数据的一致性。此外，68.4% 的标注序列跨越两个以上情绪象限；八个视频在效价维度的标注极差值范围是 $[2.475, 6.157](M = 4.637, SD = 0.859)$，唤醒维度的标注极差值范围是 $[2.678, 6.532](M = 4.831, SD = 1.074)$，这表明尽管视

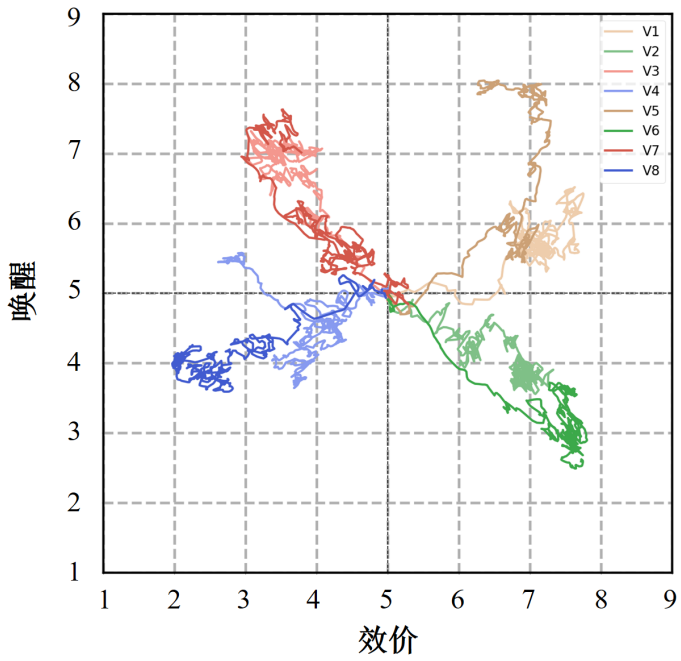


图 4.3　CEAP-360VR 数据集实时连续情绪标注数据的轨迹图

频的情绪标签唯一，但用户的实时标注数据并没有局限在此类型。

32 位被试观看八个视频的唤醒-效价情绪标注均值如图4.4所示，可以发现八个视频的唤醒和效价标注均值与诱发素材标签（见表3.2）一致。例如，V1 和 V5 属于 HVHA 类型，唤醒和效价的标注均值大于 5。为了进一步测试不同视频之间的标注差异，进行推论统计分析。W 检验表明唤醒和效价的均值都不符合正态分布（$p < 0.05$），选择 Friedman 秩和检验方法分析八个视频之间的情绪标注差异，结果为：效价（$\chi^2(7) = 146.44, p < 0.01$），唤醒（$\chi^2(7) = 120.48, p < 0.01$），表明不同视频的唤醒和效价标注结果存在显著性差异。为了明确任意两个视频在唤醒-效价两个维度的情绪标注之间是否存在差异，采用 Bonferroni 成对比较方法进行事后检验。结果如图4.5所示，在效价维度上具有显著性差异的两个视频的效应值范围是 $[0.943, 1.675]$，在唤醒维度上的效应值范围是 $[0.815, 1.655]$。结果表明，大多数情况下，情绪标签相同（如：唤醒值标签均为“高”）的两个视频标注结果没有显著性差异（$p > 0.05$），情绪标签相反的视频的标注结果具有高显著性差异（$p < 0.001$）；但在个别情况下，与假设有所差异。

对于效价维度的标注结果，标签为高效价值（HV）的视频（V1、V2、V5 与 V6）和低效价值（LV）视频（V3、V4、V7 与 V8）之间存在高显著性差异（$p < 0.001$）；但 V4 和 V8、V7 和 V8 之间也分别存在显著性差异（$0.001 < p < 0.05$），原因可能是 V8 描

述的是地震后的余震场景，用户沉浸其中后相比较其他视频会产生更低的效价值。对于唤醒维度的标注结果，标签为高唤醒值（HA）的视频（V3、V5 与 V7）和低唤醒值（LA）视频（V2、V4、V6 与 V8）之间存在高显著性差异（$p < 0.001$）；但视频 V1（镜头位于一群宠物狗中间）和 V2、V4 与 V8 之间没有高显著性差异（$0.001 < p < 0.05$），超过一半的用户在访谈环节表示非常喜爱宠物狗，因此在观看视频 V1 时的状态更放松。视频 V4 和 V6 的唤醒值标注也存在显著性差异（$p < 0.05$），V6 描述了夏威夷日出场景，用户在观看过程中的唤醒值评分非常低。

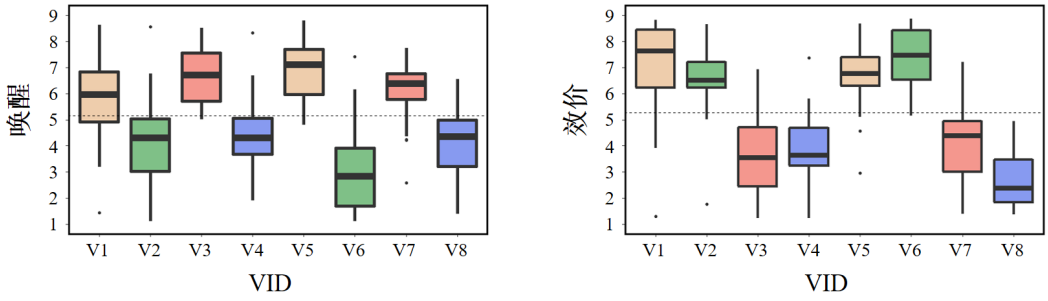


图 4.4　CEAP-360VR 数据集实时连续情绪标注数据的均值结果箱线图

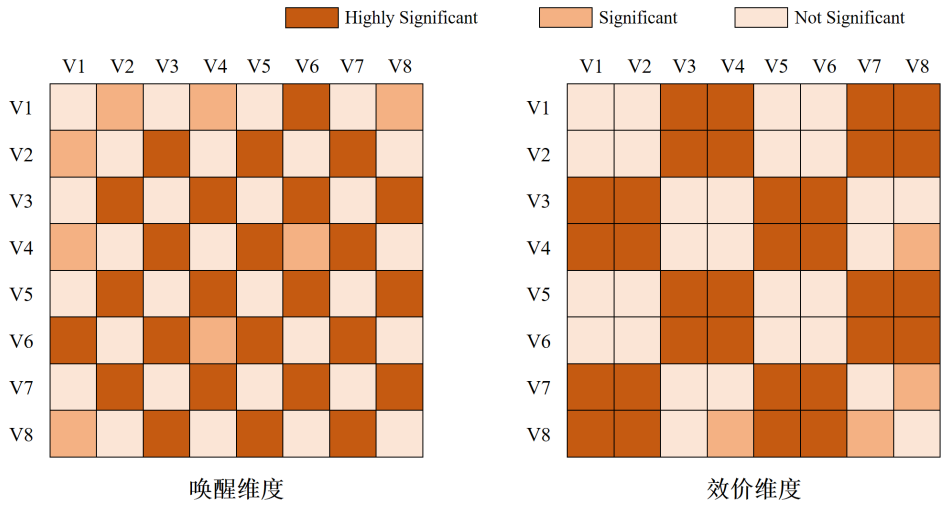


图 4.5　CEAP-360VR 数据集实时连续情绪标注数据的成对比较对称矩阵图

4.4.2　瞳孔直径数据结果与讨论

研究表明人眼的瞳孔直能够反应唤醒维度的情绪状态[152,184]，但该指标还会受到环境光强度的影响[168,185]。基于此，Pflegingde 等人[167] 与 Tarnowski 等人[67] 将用户的 PD 值建模为两个相关要素之和：（1）指定环境光照条件下的 PD 值；（2）指定体验任

务中的 PD 值。在本实验中，考虑到用户在虚拟环境中的视觉区域完全被 HMD 覆盖，除 HMD 呈现的内容之外没有其他光源信息，因此对于每个被试 $p \in [1, I]$，因观看视频 $v \in [1, J]$（I 和 J 分别表示被试总数量和诱发素材总数量）而诱发的 PD 值 $PD_{p,v}$ 计算如下：

$$PD_{p,v} = PD_{p,average} - PD_{p,light} \tag{4.1}$$

其中 $PD_{p,average}$ 表示被试 p 双眼瞳孔直径原始值的均值，$PD_{p,light}$ 表示在全景视频 v 亮度条件下用户 p 的 PD 值。$PD_{p,light}$ 的计算根据 Tarnowski 等人[67] 针对瞳孔直径在情绪识别中的研究，采用线性回归方法（系数为 k, b）对被试 p 观看视频 v 时的 PD 值与视频 v 环境光亮度之间的相关性建模如下:

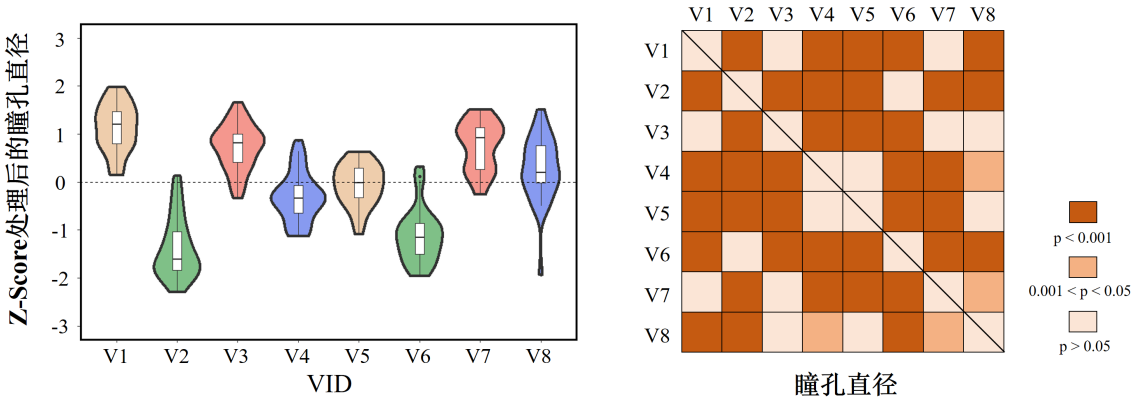
$$\begin{bmatrix} PD_{p,1} \\ PD_{p,2} \\ \vdots \\ PD_{p,n} \end{bmatrix} = \begin{bmatrix} Light_{v,1} \\ Light_{v,2} \\ \vdots \\ Light_{v,n} \end{bmatrix} * \begin{bmatrix} k \\ b \end{bmatrix} \tag{4.2}$$

其中 n 表示视频 v 的帧总数量，PD 表示被试 p 体验视频 v 时双眼原始 PD 值在每一帧的均值，$Light$ 表示视频 v 每一帧图像采用 HSV 颜色空间中 V 分量计算得到的亮度值。被试 p 受视频 v 环境光亮度影响的 PD 值计算如下：

$$PD_{p,est} = k_p * Light_v + b_p \tag{4.3}$$

其中 $PD_{p,est}$ 用作公式4.1中 $PD_{p,light}$ 的预估值。

本实验采用上述方法获取的受视频内容影响的用户 PD 值 $PD_{p,v}$，分析不同情绪类型视频中用户瞳孔直径的分布情况。首先计算所有用户在观看每个视频时 $PD_{p,v}$ 的均值与标准差，采用 Z-Score 标准化方法去除不同被试之间的基线差异。八个视频中用户受视频内容影响的瞳孔直径均值分布结果呈现在图4.6(a)中。W 检验表明 PD 数据不符合正态分布（$p < 0.001$），因此选择 Friedman 秩和检验方法进一步分析不同情绪类型视频之间的 PD 值差异。检验结果为：$\chi^2(7) = 155.98, p < 0.001$，这表明视频的情绪类型对于 PD 值具有显著影响。为了明确用户在观看任意两种情绪类型视频时的 PD 值之间是否存在差异，采用 Bonferroni 成对比较方法进行事后检验，结果如图4.6(b)所示，具有显著性差异的两个视频的效应值范围是 $[0.475, 0.859]$。实验结果表



(a) 受视频内容影响的瞳孔直径均值小提琴图　　(b) 受视频内容影响的瞳孔直径成对比较矩阵图

图 4.6　CEAP-360VR 数据集瞳孔直径数据的均值结果

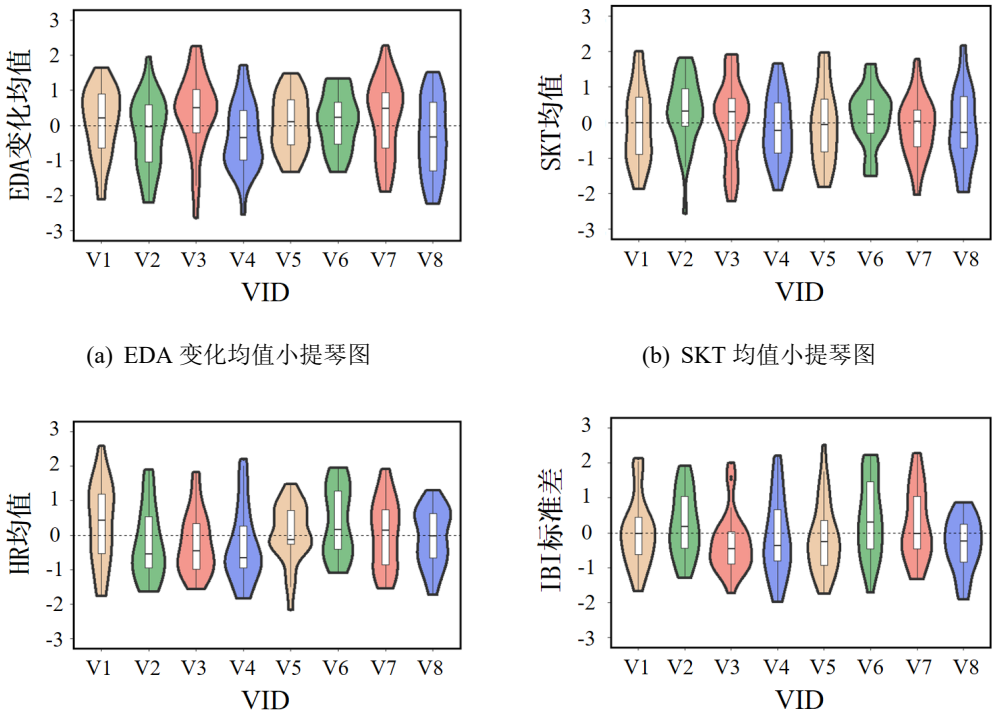
明，情绪唤醒值对瞳孔直径的影响非常显著。例如，标签为高唤醒值（HA）的视频（V1，V3，V7）与低唤醒值（LA）视频（V2，V4，V6）之间具有显著性差异；值得注意的是，标签为低唤醒值-低效价值（LALV）视频中这一结果并不明显，这表明对于“悲伤”类型的视频，用户的情绪唤醒值要比“轻松”类型（LAHV）的视频高。

4.4.3　外周生理信号结果与讨论

在生理信号的研究中[186]，一个常见问题是信号噪音对生理特征稳定性和准确性的影响。本实验选择从 Empatica E4 手环中提取的 EDA、SKT、HR 和 IBI 四种外周生理信号，分析这些信号特征与情绪之间相关性。首先参照 Nabian 等人[187] 的研究，采用截断频率为 2Hz 的三阶低通滤波器去除这些生理信号的伪影，然后将过滤后的 EDA、SKT、HR 生理信号序列正则化至 $[0,1]$，消除个体差异对生理信号的影响：

$$Normalize_{signal}(i) = \frac{signal(i) - signal_{min}}{signal_{max} - signal_{min}} \tag{4.4}$$

为了分析所有用户在观看不同情绪类型视频时生理信号的分布情况，从被试的 EDA、SKT、HR、IBI 四种外周生理信号中分别提取一个主要特征。对于皮肤电活动，计算 EDA 信号一阶导数均值作为 EDA 变化值特征[79,163]；对于皮肤温度和心率信号，分别采用 SKT 均值和 HR 均值描述信号的时域变化[159]；对于心跳时间间隔，提取 IBI 时长的标准差记为 IBI 变化值特征。由于在数据采集环节被试 $P2$ 和 $P12$ 的 IBI 数据缺失，本实验中对其余 30 位被试的 EDA 变化值、SKT 均值、HR 均值和 IBI 变化值



(a) EDA 变化均值小提琴图　　(b) SKT 均值小提琴图

(c) HR 均值小提琴图　　(d) IBI 标准差小提琴图

图 4.7　CEAP-360VR 数据集外周生理信号的均值结果

四个外周生理信号特征进行 Z-Score 标准化处理，30 位被试在观看八个视频中的上述四个生理特征分布情况呈现在图4.7中。实验结果与 Sharma 等人[78] 的研究类似：不同情绪类型视频之间的用户生理信号特征没有显著性差异。一个可能的原因是本实验中诱发素材时长设置为一分钟，这一时长对于明显的皮肤电活动较短[2]；但选取时长不一致的诱发素材难以进行标准化分析[76]，且时长较长的全景视频会给用户带来晕动症和较大的认知负荷[63,154]，也会对生理指标产生影响，因此需要在诱发素材时长与生理信号质量之间作出权衡。另一方面，图4.7的结果表明，部分生理信号特征能够表征特定情绪类型的视频。Fleureau 等人[163] 与 Boucsein 等人[188] 的研究指出，EDA 信号可以衡量皮肤电导的局部变化，与用户的唤醒值高度相关。本实验中视频 V3 和 V7 的情绪标签为高唤醒值（HA），相比于其他视频诱发了用户较高的 EDA 变化，这与研究[163,188] 的结果一致；视频 V4 和 V8 的情绪标签为低唤醒值/低效价值（LALV），用户的四个生理特征值（EDA 变化均值、SKT 均值、HR 均值和 IBI 标准差）均低于其他视频。

[2]https://support.empatica.com/hc/en-us/articles/360030048131-E4-data-EDA-Expected-signal

从 Empatica E4 手环中提取的原始生理信号序列还可用于构建用户独立和用户相关的情绪识别模型[17]；生理信号特征是随时间变化的连续序列，用户的峰值与谷值和体验事件之间密切相关，可进一步研究生理信号和连续情绪标注数据与观看行为之间的相关性。

4.5 用户数据基线验证

为了进一步验证 CEAP-360VR 数据集的有效性和可信度，本节采用常见的机器学习技术进行一系列情绪分类基线实验。

4.5.1 实验设置

本实验共设置了三项分类任务[42]：（1）二分类，将唤醒和效价两个情绪维度的标签值分为低（L）、高（H）两个类别；（2）三分类，将唤醒和效价两个维度的标签值分为低（L）、中（N）、高（H）三个类别；（3）五分类，唤醒-效价二维情绪空间的四个象限分别对应四种情绪类别（HH、HL、LH、LL），以及中立类别（NN）。实时连续情绪标注的唤醒值、效价值和离散类别之间的对应关系见表4.2。

表 4.2 连续情绪标注值与离散情绪类别之间的映射关系

类别	唤醒/效价标注值（二分类）	唤醒/效价标注值（三分类）
低（L）	[1, 5)	[1, 3)
中（N）	\	[3, 6)
高（H）	[5, 9]	[6, 9]
五分类	**效价标注值**	**唤醒标注值**
高-高（HH）	[5, 9]	(5, 9]
高-低（HL）	(5, 9]	[1, 5]
低-低（LL）	[1, 5]	[1, 5)
低-高（LH）	[1, 5)	[5, 9]
中-中（NN）	5	5

实验采用传统的机器学习方法与深度学习方法，对唤醒-效价两个维度情绪标签进行分类和预测[42]。对于机器学习方法，本实验测试了 SVM[97]、RF[98]、GaussianNB[99]、

K-NN[100] 四种分类算法；对于深度学习方法，本实验测试了情感计算中最基础且最常用的两种方法[169]：一维卷积神经网络（1D-CNN）[189] 和 LSTM[101]。

4.5.1.1 特征与模型选择

将所有被试观看八个视频的多模态用户数据划分为固定时长的片段（片段时长的设定见4.5.1.3小节），从每个片段数据中提取特征用于机器学习分类实验。本实验中选取的特征值见表4.3，分为头部、眼部行为特征与瞳孔直径特征和外周生理信号特征两大类，均为情绪识别任务中常见的行为和生理信号特征[17,38,190]。除了随机森

表 4.3 分类实验特征值

头部/眼部行为和瞳孔直径特征		外周生理信号特征	
头部运动俯仰角均值	眼部注视点时长标准差	EDA 均值	BVP 一阶导数均值
头部运动俯仰角中值	眼部注视点时长最大值	EDA 中值	BVP 一阶导数中值
头部运动俯仰角标准差	眼部注视点时长最小值	EDA 标准差	BVP 一阶导数标准差
头部运动偏航角均值	眼跳时长均值	EDA 一阶导数均值	BVP 二阶导数均值
头部运动偏航角中值	眼跳时长标准差	EDA 一阶导数中值	BVP 二阶导数中值
头部运动偏航角标准差	眼跳时长最大值	EDA 一阶导数标准差	BVP 二阶导数标准差
眼部运动俯仰角均值	眼跳时长最小值	EDA 二阶导数均值	HR 均值
眼部运动俯仰角中值	眼跳振幅均值	EDA 二阶导数中值	HR 中值
眼部运动俯仰角标准差	眼跳振幅标准差	EDA 二阶导数标准差	HR 标准差
眼部运动偏航角均值	眼跳振幅最大值	BVP 均值	SKT 均值
眼部运动偏航角中值	眼跳振幅最小值	BVP 中值	SKT 中值
眼部运动偏航角标准差	情绪相关瞳孔直径均值	BVP 标准差	SKT 标准差
眼部注视点数量	情绪相关瞳孔直径中值		
眼部注视点时长均值	情绪相关瞳孔直径标准差		

林方法，其余机器学习分类器均选择默认参数。当采用用户依赖评估模型时，RF 分类器保持默认参数（最大深度为 2）；当采用用户独立评估模型时，考虑到训练集的数量和复杂度更高，增加树最大深度（最大深度为 4）以更好地学习特征和标签之间的潜在关系。所有机器学习模型均在显卡为 NVIDIA RTX2080Ti 的机器上完成训练和测试。对于深度学习方法，采用一维卷积神经网络和长短记忆模型对 CEAP-360VR

数据集中的行为和生理数据进行分类预测。其中 1D-CNN 模型由一维卷积层、一维池化层和全连接层三部分构成。卷积层用于挖掘多模态用户数据中的深层特征，本实验中的 1D-CNN 模型包含五个一维卷积层，每层卷积核的数量 n 和长度 s 分别为（4,64）、（16，32）、（64,16）、（128,8）、（128,32），所有卷积层均采用修正后的线性激活函数（Rectified Linear Activation Function，ReLU）进行激活；一维全局最大池化层用于从卷积层中选择最显著的特征；全连接层的功能是根据选取的显著特征实现数据分类，本实验中分类函数采用 Softmax 激活函数。LSTM 模型由一个单元数为 100 的 LSTM 层构成，用于分类的全连接层与 1D-CNN 模型一致。本实验中两个深度学习网络使用 Keras 搭建在 Windows10 操作系统平台上，采用 RMSprop 优化器[191]；训练（50 epochs）和验证在显卡为 NVIDIA RTX2080Ti 的服务器上完成。

4.5.1.2 评估指标与评估方法

为了评估不同方法在不同分类任务中的分类性能，本实验选择如下两个验证指标：（1）准确率（Accuracy，acc），预测正确的百分比；（2）加权 F1-评分（Weighted F1-Score，w-f1），每个标签精确率和召回率的调和平均值。这两个指标广泛用于评估机器学习计算法[192]，考虑到标签的不平衡问题，本实验中采用加权 F1-评分代替宏 F1-评分（Macro F1-Score）和二进制 F1-评分（Binary F1-Score）。

实验选择用户依赖（SD）和用户独立（SI）两种模型对所有分类方法进行训练和评估。为了评估分类器的泛化性，对所有被试的行为数据和生理信号数据进行分组，大部分数据作为训练集用于模型训练，少部分数据作为测试集进行模型评估。对于 SD 模型，采用十折交叉验证法（10-fold Cross Validation，10-fold CV）评估，将每个被试的数据分为 10 等份，依次将其中九份作为训练数据、一份作为测试数据，用于模型训练和评估实验。对于 SI 模型，采用留一法交叉验证（Leave-One-Subject-Out Cross Validation，LOSOCV），该方法是情绪识别的标准验证方法，可用于测试分类器在不同被试之间的泛化性；选择一位被试的数据作为测试数据，其余所有被试的数据作为训练数据，上述测试和训练流程重复 N 次（N 为被试总数量）以确保所有被试数据均被用作测试数据。本实验中展示的结果为测试数据中每折或是每个被试的 acc 均值和 w-f1 均值。

4.5.1.3 片段时长选择

细粒度情绪识别需要将连续信号分割为更小（粒度更细）的片段，并识别片段中用户的情绪状态[42]。片段时长是情绪识别算法中需要重点考虑的关键参数，过长的片段中可能会包含多种情绪，导致分类结果不精确；长度越短，可以识别的情绪粒度越精细，但由于用户的情绪状态是基于每个片段中的用户多模态数据进行分类，越短的片段可能没有足够信息完成分类任务。Paul 等人[193] 指出人类一种情绪的持续时长通常在 0.5-4 秒之间，Zhang 等人[42] 最近的研究发现 1-4 秒时长的片段可作为细粒度情绪标签用于情绪分类。

为了选择适当的信号片段长度，本实验依次将多模态用户数据划分为 1s、2s、3s 和 4s 时长的片段，分别执行机器学习和深度学习实验，结果表明所有分类方法的分类结果准确度之间差异很小。先前的研究[42] 指出 2s 时长的片段足够用于细粒度情绪识别，因此在本研究后续实验中仅关注 2s 时长的片段分类结果。CEAP-360VR 数据集中列出了 1-4 秒时长片段所有分类器的分类结果，其中 RF 分类器的分类结果见表4.4。

表 4.4 RF 分类器针对不同时长片段的基线验证实验结果

验证模型	片段时长	Valence-2		Arousal-2		Valence-3		Arousal-3		5 - Class	
	(s)	acc	f1	acc	f1	acc	f1	acc	f1	acc	f1
SD (10-fold CV)	1	67.83%	0.6194	71.50%	0.6459	57.61%	0.5069	60.90%	0.5287	49.04%	0.4095
	2	68.45%	0.6315	71.33%	0.6487	60.42%	0.5354	62.38%	0.5457	51.89%	0.4340
	3	69.70%	0.6443	71.08%	0.6459	62.16%	0.5566	64.31%	0.5683	53.99%	0.4557
	4	69.78%	0.6482	71.80%	0.6599	62.43%	0.5599	64.96%	0.5764	56.34%	0.4762
SI (LOSOCV)	1	64.90%	0.5378	65.94%	0.5307	46.30%	0.4049	51.84%	0.4176	33.35%	0.2226
	2	66.80%	0.6238	64.26%	0.5298	49.92%	0.4419	52.20%	0.4341	31.47%	0.3001
	3	65.49%	0.5870	63.28%	0.4973	48.54%	0.4216	51.76%	0.4154	33.34%	0.3078
	4	65.73%	0.5981	62.40%	0.4864	50.44%	0.4413	51.77%	0.4163	34.38%	0.3002

4.5.2 分类实验结果与讨论

在四种机器学习方法的分类实验中，RF 分类器的分类结果准确度略优于 SVM、NB 和 K-NN 分类器，因此本文仅展示了 RF 分类器的实验结果并将其用于后续分析讨论。其余所有分类器的实验结果呈现在 CEAP-360VR 数据集中。

本节研究了 RF、1D-CNN 与 LSTM 三种分类方法在 2 秒时长片段下、采用用户依赖和用户独立两种评估模型的分类性能，实验结果见表4.5。CEAP-360VR 数据集三分类任务的准确度低于二分类，但高于五分类。值得注意的是，数据集中很多片段的情绪标签为“中立”，其中三分类任务中情绪标签为“中（N）”的片段比例为 43.32%，五分类中标签为“中-中（N-N）”的片段比例为 27.06%，数据类型的分布不均衡会影响细粒度情绪识别的准确度。表4.5中序列学习方法 LSTM 在细粒度情绪识别中并没有取得较好的效果，由于情绪标签类型分布不均衡，较多的“中立”情绪标签容易造成序列学习模型的循环结构过拟合。与之类似，Zhang 等人[42] 的一项研究分析了电脑端（CASE）[78] 和移动端（MERCA）[79] 两个数据集的实时连续标注数据，研究指出 CASE 数据集中超过 60% 的片段情绪标签为“中立”类别，MERCA 数据集中超过 50% 的片段情绪标签为“中立”；并指出这一数据不平衡现象是因为持续标注过程中被试倾向于在默认情况下释放摇杆设备，从而将其标注为“中立”类别。在未来的研究中，可以采用 CEAP-360VR 数据集中的实时连续情绪数据训练生成模型（Generative Adversarial Networks，GAN）[194]，通过人工生成样本扩大特定情绪类型的数据规模。

此外，在表4.5中，相比于用户独立评估模型，用户依赖模型的 acc 值和 w-f1 值更高，特别是在三分类和五分类中。这一结果表明单个用户的数据量足够用于训练机器学习模型进行情绪识别，同时也验证了本实验中单个用户观看的视频数量和视频时长足够用于执行分类实验。综上，典型的机器学习和深度学习方法能够用于对 CEAP-360VR 数据集中的多模态用户数据执行情绪分类任务。

表 4.5　RF、1D-CNN 和 LSTM 分类器针对 2s 时长片段的分类实验结果

验证模型	分类器	Valence-2		Arousal-2		Valence-3		Arousal-3		5 - Class	
		acc	f1	acc	f1	acc	f1	acc	f1	acc	f1
SD (10-fold CV)	RF	68.45%	0.6315	71.33%	0.6487	**60.42%**	**0.5354**	**62.38%**	**0.5457**	**51.89%**	**0.4340**
	1DCNN	**68.46%**	**0.5739**	71.37%	0.6121	51.17%	0.3867	56.85%	0.4502	40.49%	0.2828
	LSTM	65.51%	0.6013	**71.39%**	**0.6560**	52.91%	0.4755	56.36%	0.5002	44.48%	0.3735
SI (LOSOCV)	RF	**66.80%**	**0.6238**	64.26%	0.5298	**49.92%**	**0.4419**	**52.20%**	**0.4341**	**31.47%**	**0.3001**
	1DCNN	64.27%	0.5828	**67.64%**	**0.5808**	45.51%	0.4191	47.17%	0.3923	29.87%	0.2598
	LSTM	65.00%	0.6349	66.30%	0.5934	44.62%	0.4269	43.79%	0.4085	30.12%	0.2768

4.5.3 消融实验结果与讨论

为了进一步分析 CEAP-360VR 数据集中每种模态数据的有效性，本节执行消融实验，用于查看行为数据（头部和眼部运动、瞳孔直径）和外周生理信号（EDA、BVP、HR、SKT）对情绪分类结果的影响。表4.6中列出了 RF 分类器在 2 秒时长片段下、采用用户依赖和用户独立两种评估模型的三种分类任务结果。结果表明仅采用行为数据或是外周生理信号数据能够产生较好的识别准确度；将行为数据和生理信号相结合相比单一模态类型的识别准确度更高。

表 4.6　行为数据与外周生理信号的消融实验结果

验证模型	元素	Valence-2		Arousal-2		Valence-3		Arousal-3		5 - Class	
		acc	f1	acc	f1	acc	f1	acc	f1	acc	f1
SD (10-fold CV)	**Physio Only**	65.9%	0.608	68.3%	0.612	54.8%	0.479	60.0%	0.524	45.2%	0.365
	HM/EM + PD	67.4%	0.621	69.2%	0.627	58.3%	0.515	61.5%	0.533	50.6%	0.422
	Physio + HM/EM + PD	**68.5%**	**0.632**	**71.3%**	**0.649**	**60.4%**	**0.535**	**62.4%**	**0.546**	**51.9%**	**0.434**
SI (LOSOCV)	**Physio Only**	65.7%	0.613	62.2%	0.501	44.7%	0.403	51.5%	0.420	30.6%	0.244
	HM/EM + PD	62.7%	0.534	62.4%	0.502	47.9%	0.412	50.0%	0.356	26.3%	0.244
	Physio + HM/EM + PD	**66.8%**	**0.624**	**64.3%**	**0.530**	**49.9%**	**0.442**	**52.2%**	**0.434**	**31.5%**	**0.300**

4.6 本章小结

本章构建了虚拟交互环境中首个公开的生理及行为多模态连续情绪数据集 CEAP-360VR，包含 32 位被试观看八个全景视频时的情绪标注数据、头部与眼部运动数据、瞳孔直径数据和外周生理信号数据（EDA、IBI、HR、SKT、BVP），以及多模态数据的预处理脚本和统计学分析及基线实验脚本文件。描述性统计学和基线实验的结果表明：（1）CEAP-360VR 数据集中连续唤醒-效价标注均值与诱发素材的情绪标签一致，表明实时连续情绪标注数据的可信度[78]；（2）瞳孔直径数据与唤醒维度标注数据具有正相关，同时视频的环境光对唤醒维度标注也有一定影响，这与先前的研究[167,168]一致；（3）基线实验表明 RF 分类器在 2 秒时长片段中的分类性能较好，对于 SD 评估模型，二分类任务的效价和唤醒准确率分别是 68.45%、71.33%，三分类任务的效价和唤醒准确率分别是 60.42%、62.38%，对于 SI 模型，二分类任务的效价和唤醒维度准确率分别是 66.80%、64.26%，三分类任务的效价和唤醒维度准确率分别是 49.92%、

52.20%；（4）消融实验结果表明仅使用行为数据或是生理信号能够产生合理的识别准确度，但同时使用这两类模态信息能够提升识别准确度。

CEAP-360VR 数据集是虚拟环境中第一个公开的多模态情绪数据集，同时具有实时连续情绪标注数据、行为和生理信号数据。该数据集一方面可用于开发和验证细粒度情绪识别算法，构建更精确的时序性情绪识别模型；另一方面，多模态用户数据能够探索头部与眼部运动和离散与连续情绪标签之间的关系，多样的生理信号数据可以开展隐式感知体验分析。综上，CEAP-360VR 数据集可以进一步促进情感计算研究领域对沉浸式虚拟环境中连续情绪状态及其与生理和行为反应的理解。

第 5 章　视觉交互行为与情绪的相关性识别

5.1　引言

在虚拟环境中，用户通过头部运动实时选择视口位置和观看内容，体验过程更具有交互性。这一交互行为带来了更好的沉浸感与临场感[28]，但自由多样的视觉行为也为用户体验质量评估提出了新的挑战。鉴于此，视觉注意力（Visual Attention，VA）[195]研究在虚拟现实和多媒体领域迅速发展，通过改进沉浸式交互媒体的处理、编码、传输和渲染技术，提供一种低成本方法提升用户的体验质量。如何获取准确的头部运动等视觉交互行为、理解用户在虚拟环境中的视觉特征和探索模式至关重要。

情绪在用户体验质量评估中起着关键作用。研究[18,196,197]表明人的头部姿态和运动能够反映内在的情绪状态，例如，人在开心时往往会抬起头，而在心情低落时会低下头。同时，人的眼部运动受到大脑的影响，在情绪刺激下眼球运动也会发生变化[198,199]。但是关于虚拟环境中用户头部行为与眼部运动和情绪状态之间潜在联系的研究非常稀缺。最近，Li 等人[63]研究了用户在观看全景视频时头部运动与情绪之间的关系，如图5.1所示；Tang 等人[38]分析了用户在虚拟环境中观看全景图片时情绪对于眼部行为的影响，但是这些研究中的情绪标签均来自体验后标注。由于情绪在诱发素材作用下实时产生且连续变化，需要进一步在更细粒度的层级上探索实时连续情绪状态和视觉交互行为之间的相关性。

本章围绕头部运动与眼部运动，研究虚拟体验中用户的视觉交互行为特征、及其与连续情绪数据之间的相关性。主要内容分为以下两个方面：

（1）介绍了头部运动与眼部运动在真实环境、虚拟空间及等距投影三种情境下的转换方式，计算头部运动扫描路径并从眼部运动中提取注视与眼跳两个眼动行为特征；基于视觉注意相关研究，探讨了虚拟环境中用户的四种视觉交互行为特征：用户之间的头部运动与眼部运动一致性、用户头部运动与眼部运动之间的相关性、用户视觉行为的赤道及前方偏向、诱发素材内容对用户视觉行为的影响。

（2）提出了一种细粒度层级的视觉交互行为与连续情绪标签之间相关性识别方法；通过计算用户在虚拟体验中头部与眼部运动数据和实时连续情绪标注数据之间的皮尔逊积矩相关系数，对不同时长片段中视觉行为与连续情绪标签之间的相关性进行评估；研究还探索了实时连续情绪标签和眼部运动的注视与眼跳特征之间的相关性。

实验结果表明，CEAP-360VR 数据集中 32 位被试观看八个全景视频的头部运动与眼部运动均具有很高的用户间一致性；被试视觉注意会受到观看内容影响，但仍能观察到明显的赤道和前方偏向。用户观看行为与情绪标注相关性的实验结果表明：对于 5s、10s 和 20s 时长的片段，头部运动偏航角的标准差与情绪效价值呈负相关，而头部运动俯仰角与唤醒值呈正相关；对于 5s 和 10s 时长的片段，眼部运动偏航角标准差与效价值呈负相关，而眼部运动俯仰角与唤醒值呈负相关；用户观看“HVHA”类型视频时眼部注视数量多而眼跳时长短。以下将会详细介绍方法和实验结论。

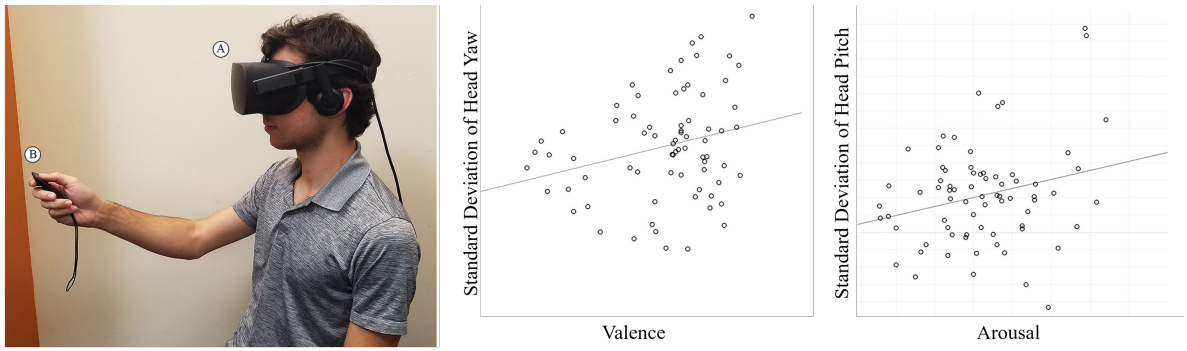


图 5.1　头部运动与体验后情绪标注的关系研究[63]

5.2　相关工作

在虚拟现实研究领域，很多公开的数据集报告了用户的头部运动轨迹与眼部运动数据，主要用于 VAM 研究。Ozcinar 等人[200] 采集了 17 名被试观看六个全景视频时的视口中心点位置，发现用户视觉注意会受到移动物体的引导；Xu 等人[201] 构建的 PVS-HM 数据集表明所有被试的观看模式均具有很高的相似性且关注区域集中在视频中心附近。Rai 等人[202] 收集了 63 名被试观看全景图片时的头部运动和眼部运动数据，随后 David 等人[179] 将其扩展至全景视频中，提出的 Salient360 数据集包可用于虚拟环境中媒体内容的观看行为模式与视觉注意研究。为了进一步探索沉浸式全景视频观看中用户头部运动与眼部运动显著性预测模型，Xu 等人[203] 提出了一个包含 31 名被试观看 208 个视频时的头部运动和眼部运动的大规模 VR 数据集；Nguyen 等人[204] 构建了一个显著性数据集并提出了 PanoSalNet 显著性检测模型。

视觉交互行为数据除了用于 VAM 研究，还可用于探索用户的情绪状态。研究[18,205,206] 表明头部运动相比于面部表情能够传递更多的重要信息。Lhommet 等人[207] 和 Gross 等人[208] 的研究指出特定的头部运动和特定的情绪之间存在显著相关性。Livingstone

等人[178]记录了声乐家在演唱不同情绪段落时的头部运动，结果表明头部运动的俯仰角与情绪之间具有显著相关性。Lemos 等人[198]提出眼部运动的凝视特征，包括眨眼和瞳孔变化，能够有效预测唤醒和效价两种情绪。Wiebe 等人[209]的研究发现用户观看积极或是消极情绪的图片时会比中性情绪图片花费更多时间。但是，这些研究均是在非沉浸式环境中进行。

在虚拟环境中，Slater 等人[176]对 20 名用户开展了一项实验，用户在虚拟场地中漫游并计算带有患病树叶的树木数量，结果显示头部运动偏航角与用户报告的临场感之间具有显著正相关。Won 等人[177]的研究指出虚拟体验中用户的侧向头部转动和焦虑值之间存在相关性。Li 等人[63]分析了用户观看全景视频时头部运动和唤醒与效价值的相关性，研究发现头部运动的俯仰角与唤醒值间具有显著正相关，偏航角的标准差与效价值间呈现正相关，如图5.1所示。Tang 等人[38]初步探索了用户在观看全景图片时情绪状态对眼部行为特征的影响，并提出消极情绪对于眼部的注视数据和眼跳特征具有显著影响。

上述研究表明，虚拟环境中用户的头部运动与眼部运动和情绪状态之间存在相关性，能够用于推断用户情绪状态。但是，现有研究均没有收集用户实时连续的情绪报告标签，也没有在细粒度层级上分析情绪状态和行为特征之间的关系。此外，虚拟环境中用户行为与情绪的相关性研究集中在全景图片上，而全景视频[63]的研究仅采集了头部运动，没有考虑眼部运动数据。

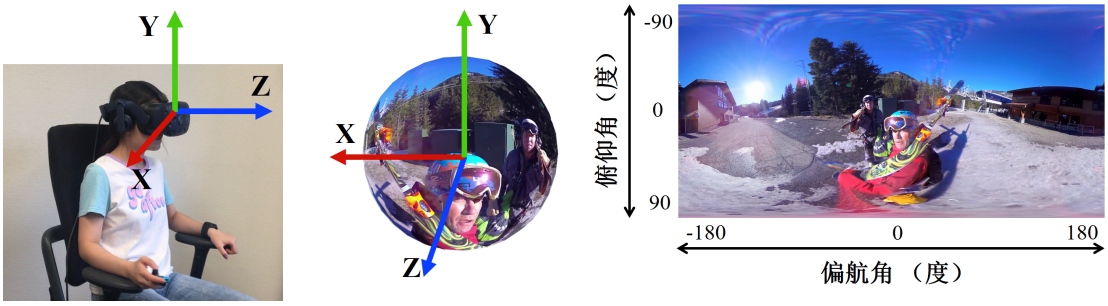
5.3 视觉交互行为

本节首先介绍了虚拟环境中用户头部运动与眼部运动的计算和分析方法；然后探讨了基于视觉注意的四种视觉交互行为特征；最后提出一种细粒度层级的视觉行为特征与实时连续情绪相关性评估方法。

5.3.1 头部运动和眼部运动

在虚拟环境中，用户可以通过头部运动（Head Movement，HM）自由地将视口聚焦在感兴趣的内容上，这一方式与现实世界一致，并为其带来了沉浸式交互体验[170]。用户佩戴 HMD 设备进行虚拟体验时的头部运动反应了视口位置信息；眼部运动（Eye Movement，EM）表示视口内用户眼睛的聚焦点[171]。David 等人[179]指出，头部运动本身无法反应用户真实的视觉聚焦点，因此虚拟环境中用户的视觉交互行为研究需要

同时考虑头部运动和眼部运动。



(a) 真实环境坐标系统　(b) 虚拟空间坐标系统　(c) 等距投影坐标系统

图 5.2　不同场景中的头部运动坐标系统标定

用户佩戴 HMD 时的头部运动和眼部运动发生在三维空间中，图5.2(a)中标注了真实空间中用户头部运动的左手坐标系。图5.2(b)所示为诱发素材投影至球体表面上的镜像效果，标注了虚拟空间中用户头部运动的左手坐标系，用户通常位于球体内部的球心位置，当用户沿 Z 轴向前看时，X 轴指向右侧、Y 轴指向上方。为了获取用户视觉行为和诱发素材内容的对应信息，从头部运动和眼部运动三维空间的原始数据中提取偏航角（Yaw）和俯仰角（Pitch）。图5.2(c)描述了诱发素材在等距圆柱投影（EquiRectangular Projection，ERP）方式下的帧画面，其中水平方向对应用户头部旋转偏航角，范围在 $[-180^o, 180^o]$，0^o 表示水平中心点；垂直方向表示用户头部运动俯仰角，范围是 $[-90^o, 90^o]$，0^o 表示等距投影的垂直中心点。对于 HTC VIVE Pro Eye HMD 设备，从 HMD 中直接获取的用户头部运动原始数据为 $rot(x, y, z)$，通过线性变换计算头部旋转的俯仰角 H_{pitch} 和偏航角 H_{yaw}，公式如下：

$$H_{pitch} = \begin{cases} rot.x & rot.x \leq 180 \\ rot.x - 360 & rot.x > 180 \end{cases} \qquad H_{yaw} = \begin{cases} rot.y & rot.y \leq 180 \\ rot.y - 360 & rot.y > 180 \end{cases} \tag{5.1}$$

HMD 中获取的眼部运动原始数据包括左眼、右眼和双眼的凝视方向，从双眼凝视方向 (x, y, z) 中提取俯仰角 E_{pitch} 和偏航角 E_{yaw} 的公式如下：

$$E_{pitch} = sgn(y)\arccos\sqrt{\frac{x^2+z^2}{x^2+y^2+z^2}}\frac{180}{\pi} \qquad E_{yaw} = \begin{cases} \arctan(\frac{x}{z})\frac{180}{\pi} & z > 0 \\ 180 + \arctan(\frac{x}{z})\frac{180}{\pi} & x > 0, z < 0 \\ -180 + \arctan(\frac{x}{z})\frac{180}{\pi} & x < 0, z < 0 \end{cases} \tag{5.2}$$

情感计算与认知研究中[38,190]常见的眼动行为特征主要有注视（Fixation）和眼跳（Saccade）。注视是指双眼在感兴趣目标上保持一定时长以获得足够注视细节的行为；而眼跳是指双眼从一个关注点移动到另一关注点的快速眼动行为。眼跳期间眼球会快速移动，在视网膜上的成像质量很差，因此多数视觉信息是通过注视行为获得的。为了计算虚拟环境中用户的眼部注视数据，首先需要计算用户在每个采样点眼部运动的速度和加速度。球体坐标系下的速度计算不能直接使用欧几里得距离，应该采用两个采样点之间的正交距离（又称大圆距离）除以它们的时间差。λ_1, ϕ_1 与 λ_2, ϕ_2 分别表示两个相邻采样点的偏航角与俯仰角值，$\Delta\delta$ 表示两个采样点之间的正交距离，计算公式如下：

$$\Delta\sigma = 2\arcsin\sqrt{\sin^2(\frac{\Delta\phi}{2}) + \cos\phi_1\cos\phi_2 + \sin^2(\frac{\Delta\lambda}{2}))} \tag{5.3}$$

将第一个采样点的速度和加速度设置为 0，其余采样点的速度和加速度通过如下方式计算：

$$v_2 = \frac{\Delta\delta}{\Delta t}, \qquad a_2 = \frac{v_2 - v_1}{\Delta t} \tag{5.4}$$

研究[210]通常采用基于阈值的方法，将眼跳行为的速度和加速度阈值分别设置为 $75^o/s$ 和 $200^o/s^2$。Salvucci 等人[211]的研究发现人眼注视发生的最短时长为 150 毫秒。假设两次眼跳之间的时间间隔为 t，当 $t > 150ms$ 时，表明用户的眼部运动在该时间段内是注视状态，该区间内所有采样点的质心即为注视点位置。

对于头部运动，用户佩戴 HMD 时快速的头部旋转会带来不适甚至严重眩晕，因此这一行为很少发生；而在低速的头部旋转中，用户可以通过头部和眼部的补偿行为（前-庭眼反射机制）自我调节，仍然能够感知内容。所以在虚拟空间中，采用头部运动的扫描路径替代“头部注视”或者“头部扫视”等概念[179]。通过将用户头部运动的数据序列划分为时长为 200 毫秒的连续窗口，计算每个窗口内所有采样点的质心作为该区间内头部运动的路径点，进而获取每个诱发素材所有被试时间对齐的头部运动

轨迹路径。

5.3.2 视觉交互行为特征

人的视觉注意机制表明，用户在观察一个场景时，能够自动聚焦感兴趣区域，而选择性地忽略其他不感兴趣区域，感兴趣区域即为显著性区域。用户的视觉注意通常采用二维显著图（Saliency Map）方法建模，特别是在全景图片或全景视频观看体验中，将全景内容从球体表面映射在二维平面上，基于用户头部运动和眼部运动采样点生成带有显著性的热图。结合 VAM 研究现状[171]，本节从用户之间与用户自身的头部运动与眼部运动关系、及视觉行为的统计学偏向两个角度出发，将虚拟环境中用户的视觉交互行为特征归纳为如下四个方面：

（1）用户之间的头部运动与眼部运动一致性。探索不同用户在同一诱发素材体验中视觉交互行为的一致性非常关键，这在一定程度上反应了视觉行为模式及视觉注意模型的泛化能力。当前主要有两种一致性分析方法。Sitzmann 等人[212] 引入了接收者操作特征曲线（Receiver Operating Characteristic Curve，ROC）作为用户观看全景图片行为一致性的度量指标，依次比较每个被试与其余所有被试视觉行为显著图之间的相似性；实验发现 ROC 曲线快速收敛至最大速率 1，表明用户的观看行为在给定场景中均具有很强的一致性。Xu 等人[213] 提出了一种显著图之间线性相关系数（Linear Correlation Coefficient，CC）的计算方法，用于测量不同用户之间观看行为的一致性，该方法将所有用户随机等分为两组，计算两组用户头部运动显著图之间的相关系数；实验发现 CC 的结果为 $M = 0.745, SD = 0.114$，表明用户观看全景视频的视觉行为有很强的相似性。

（2）用户头部运动与眼部运动之间的相关性。在虚拟环境中，用户的头部运动反应视口位置信息，眼部运动表示聚焦点，同一用户的头部运动和眼部运动可能存在差异。Rai 等人[214] 发现用户在观看全景图片时的眼部运动呈火山式分布，并定量分析了头部运动和眼部运动显著图之间的统计学差异，结果表明二者的分布类似但仍有区别。前庭-眼反射机制[58] 指出用户的眼部和头部运动相反，从而能够稳定视线提高视觉注意质量。Sitzmann 等人[212] 的研究也发现用户在注视状态下头部速度和相对凝视速度之间呈现预期反向线性相关。在视觉行为建模和预测领域中，一些研究[215,216] 采用同一模型预测头部运动和眼部运动的显著图，而更多的工作[201,217,218] 会将二者区分开，采用不同模型分别预测用户的头部运动和眼部运动视觉交互行为模式。

（3）用户视觉行为的赤道及前方偏向。研究[219]表明用户在观看传统图片或视频时，倾向于盯着屏幕或场景的中心区域，媒介的关键信息也通常呈现在中心区域，因此眼部运动存在中心偏向（Center Bias）现象[220,221]。用户在虚拟环境中通过头部旋转与内容交互时，长时间仰头或者低头均会产生不适感，因此在赤道附近的内容会更多地受到用户关注，这一现象即为“赤道偏向（Equator Bias）”[212,215]。David 等人[179]和 Xu 等人[201]的研究还指出相比于观看全景图片，用户在观看全景视频时更倾向于关注正前方区域。这意味着用户在虚拟环境中初始观察点的位置很重要，该位置信息对应“正前方”区域，可能会影响用户视觉行为模式。许多视觉注意力的研究将赤道和前方偏向作为先验知识，用于提升视觉注意预测模型的精准度[217,222]。

（4）诱发素材内容对用户视觉行为的影响。除统计学偏向之外，虚拟环境中用户的视觉注意还与诱发素材内容具有较强的相关性[213]。Sitzmann 等人[212]通过计算显著图的香农熵（Shannon entropy）描述全景内容里显著区域的分布，定量分析内容对视觉注意的影响；高熵值表示大量显著物体散布在场景中，会造成用户视觉注意的分散，低熵值表示场景中仅有个别显著物体吸引了用户所有注意。因此当环境中显著物体数量较少或位置较近时，会吸引更多的用户注意。由于在虚拟环境中用户可以通过头部运动选择观看视口区域，通过眼部运动选择具体的观看对象，分析内容对用户视觉注意的影响有助于解决预测用户观看行为的关键挑战，从而更好地分析、理解和开发虚拟交互内容素材。

5.4 视觉交互行为与情绪的相关性识别

在虚拟环境中，用户体验同一诱发素材时的视觉行为具有很高的相似性。视觉行为特征研究一方面可用于视觉注意建模，另一方面通过探索视觉行为与情绪状态之间的相关性，有助于进一步提升用户体验质量。本节借鉴 Li 等人的工作[63]，通过计算虚拟空间中用户的头部运动与眼部运动数据和实时连续情绪标签之间的皮尔逊积矩相关系数（Pearson product-moment correlation coefficient，PCCs），计为 $corr$，提出一种不同时长片段中行为特征与连续情绪之间相关性评估的方法，同时还探索了实时连续情绪标签和眼部运动的注视与眼跳特征之间的相关性。

相关性强弱采用心理学领域 Haldun 等人[223]提出的标准衡量：低，$0.1 < |corr| < 0.3$；中，$0.3 < |corr| < 0.6$；高，$0.6 < |corr| < 1.0$。值得注意的是，本方法中的多重相关性比较会产生很多假阳性结果（I 型错误）[224]，而采用 Bonferroni 方法调整过

于保守，会在降低 I 型错误的同时增加 II 型错误[225]。因此本研究中将 alpha 值从 0.05 降低至 0.01，同时还选择了更为均衡的错误发现率（False Discovery Rate，FDR）方法[226] 进行测试校准。但考虑到将 alpha 值设置为 0.01 相比于 FDR 校准的结果更为保守，实验中仅报告了降低 alpha 值的结果。此研究的目的是在细粒度层级上分析实时连续的情绪标签和头部运动与眼部运动之间的相关性。

5.5 实验结果与讨论

3.5章节的实验结果表明，CEAP-360VR 数据集（见4.3小节）中 HaloLight 与 DotSize 两种实时连续情绪标注方法之间没有显著性差异。因此，本节将两种方法收集的用户多模态数据整合，分析 32 位被试观看八个不同情绪类型视频的头部运动、眼部运动数据及其与实时连续情绪标注数据间的关系。本节主要从视觉行为特征的描述性统计分析、行为数据与情绪标注数据细粒度层级的相关性分析以及实验结果讨论三个方面逐一进行详述。

5.5.1 视觉交互行为特征结果与讨论

研究首先采用 Xu 等人[213] 提出的方法评估所有被试之间观看行为的一致性。将 CEAP-360VR 数据集中 32 位被试随机等分为两组：计为 $Group1$ 和 $Group2$，将全景视频由球面映射至二维平面上，并分别生成每一帧内两组被试头部运动和眼部运动显著图的平面坐标向量，即 H_1 和 H_2，采用线性相关系数 CC 量化二者之间的相关性，计算公式如下

$$CC(H_1, H_2) = \frac{\sum_{s,t}(H_1(s,t) - \mu(H_1)) \cdot (H_2(s,t) - \mu(H_2))}{\sqrt{\sigma(H_1)^2 \cdot \sigma(H_2)^2}} \tag{5.5}$$

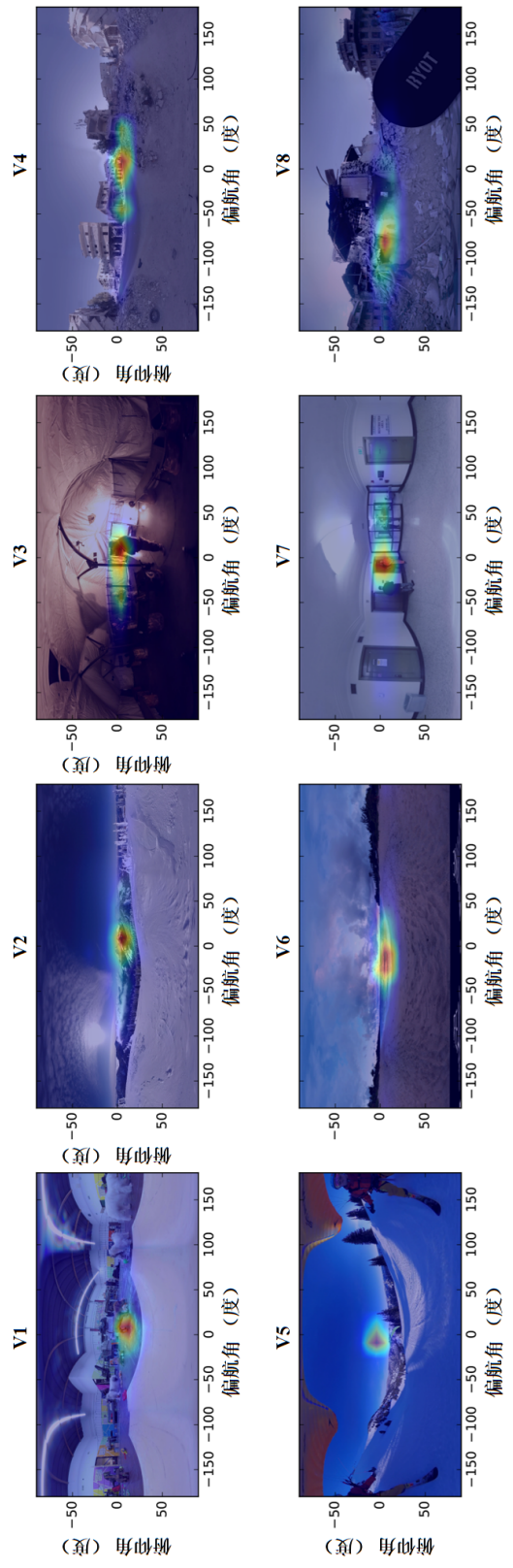
其中，(s,t) 为像素点坐标，$\mu(\cdot)$ 和 $\sigma(\cdot)$ 分别表示对应显著图的均值和标准差。CC 系数的范围是 $[-1,1]$，-1 表示完全负相关、1 表示完全正相关。表5.1中报告了两组被试头部运动及眼部运动之间相关系数的均值和标准差，结果表明 32 位被试在观看八个全景视频时的视觉行为具有较高的一致性（$CC > 0.8$）。被试观看八个视频过程中的头部运动相关系数均值为 0.877 ± 0.042，眼部运动相关系数均值为 0.952 ± 0.023，高于 Xu 等人[213] 的实验结果 0.745 ± 0.114。Shiraishi 等人[56] 的研究指出，用户在虚拟环境中通过头部运动和眼部运动进行目标选择时，更倾向于采用头部旋转。但是，从

表5.1的一致性结果可以看出，本实验中每个视频的眼部运动一致性均略高于头部运动的一致性。当被试观看相同的兴趣区域（或目标体）时，不同被试视口的具体位置略有偏差，而眼睛的注视点停留在同一个目标体上。因此，尽管先前的研究[56]表明虚拟环境中用户更倾向于通过头部运动选择观看目标，但本实验中眼部运动的一致性更高。

表 5.1　用户头部运动与眼部运动显著图的相关系数比较

VID	CC (HM)	CC (EM)
V1	0.881 ± 0.016	0.913 ± 0.012
V2	0.843 ± 0.010	0.952 ± 0.042
V3	0.862 ± 0.047	0.956 ± 0.023
V4	0.917 ± 0.050	0.967 ± 0.032
V5	0.883 ± 0.064	0.971 ± 0.013
V6	0.915 ± 0.046	0.960 ± 0.025
V7	0.861 ± 0.064	0.970 ± 0.012
V8	0.854 ± 0.042	0.926 ± 0.021
总体	0.877 ± 0.042	0.952 ± 0.023

图5.3(a)所示为八个视频中所有被试头部运动等距投影形式的显著图，图5.3(b)为眼部运动的显著图，分别由 32 位被试的头部运动和眼部运动数据采样点生成，其中 X 轴为偏航角，Y 轴表示俯仰角。为了规避初始点对用户观看行为的影响，本实验将所有被试的观看方向初始化为视频的中心点（见3.4.2（4））。从图5.3可以看出，被试的绝大多数视觉注意点集中在赤道和中心附近的区域内；除此之外，还存在一些潜在区域吸引了用户的视觉注意，这些区域与视频内容具有较强的相关性[212,213]。例如，在 V3 中，僵尸不断从视频中心及左侧区域出现，相应的被试关注区域也集中在中心和左侧；视频 V7 中被试的视角是追随嫌犯的逃跑路线，因此视觉关注的显著区域并不唯一；对于 V8，被试的主要视觉注意区域具有明显左偏，原因是整场视频中创作者在视频的右侧嵌入了黑色徽标。

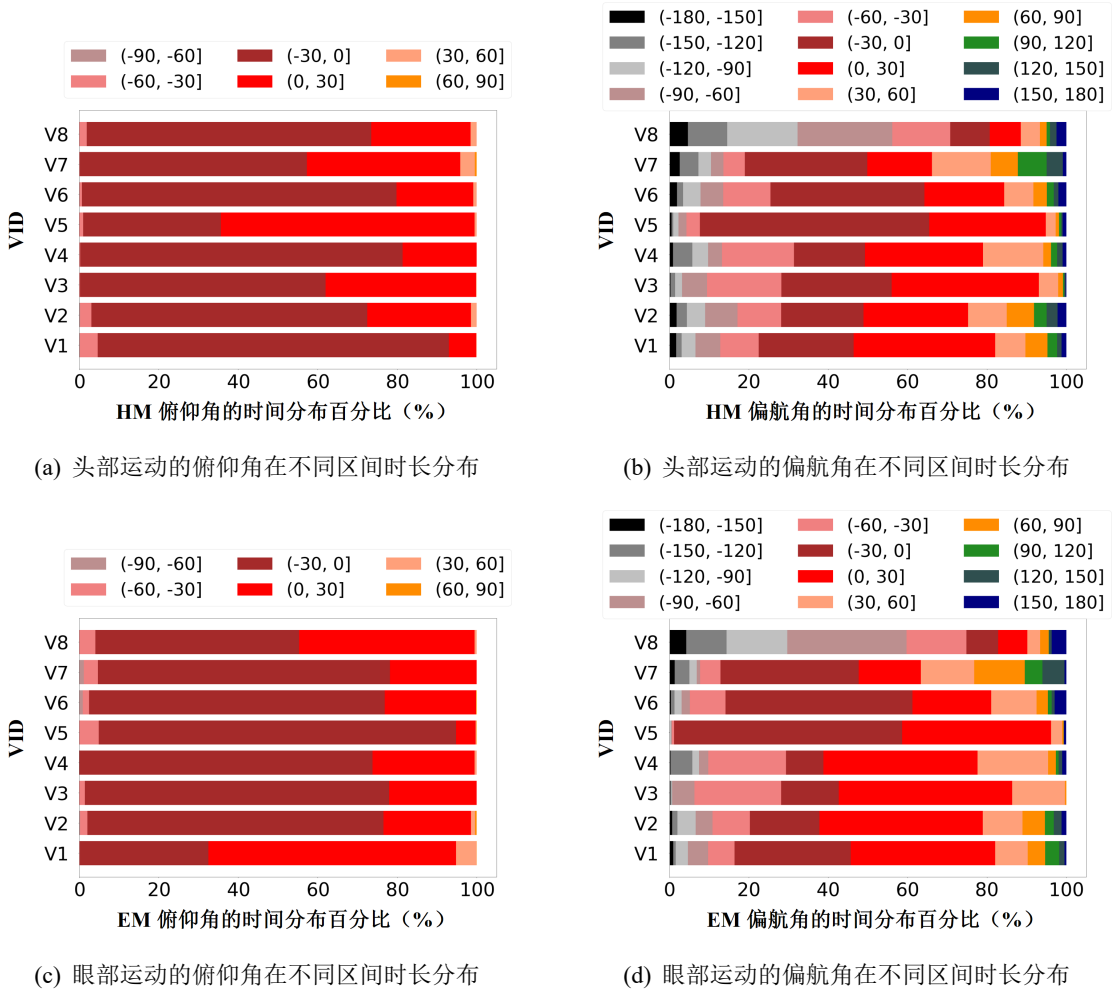


(a) 32 位被试观看八个视频时的头部运动数据显著图

(b) 32 位被试观看八个视频时的眼部运动数据显著图

图 5.3　用户头部运动和眼部运动显著图

为了进一步分析被试在观看全景视频时视觉注意的分布情况，研究计算了 CEAP-360VR 数据集中 32 位被试观看每个视频时头部运动和眼部运动所有采样点的俯仰角和偏航角的时间分布百分比，图5.4中提供了所有被试观看全景视频时在每块区域花费的时间百分比信息。头部运动和眼部运动的俯仰角分布（如图5.4(a)和图5.4(c)所示）



(a) 头部运动的俯仰角在不同区间时长分布

(b) 头部运动的偏航角在不同区间时长分布

(c) 眼部运动的俯仰角在不同区间时长分布

(d) 眼部运动的偏航角在不同区间时长分布

图 5.4　头部运动与眼部运动的俯仰角和偏航角时间分布情况

表明，被试大部分观看时间集中在视频的赤道部分，即 $-30°$ 到 $30°$ 之间，占据了 94% 的时间；就偏航角（如图5.4(b)和图5.4(d)所示）而言，视觉时间分布更多地取决于视频内容和被试的感兴趣区域。例如，在视频 V5 中，以一名飞行员的第一视角滑行飞过山脉，被试视口在 85% 以上的时间内位于视频正前方区域（$-30°$ 到 $30°$ 之间）；在 V1 中被试位于一群可爱的宠物狗中间，85% 以上的时间视口位置在 $-90°$ 到 $90°$ 之间，相比于 V5 探索区域更广。总体来看，被试头部和眼部运动数据的时间分布较为相似，但头部运动的分布范围比眼部运动更广；被试关注极端偏航区域（初始朝向正

后方）的时间很少，150° 到 −150° 区域的观看时间比例不足 8%。

5.5.2 相关性识别结果

在本实验中，首先将 CEAP-360VR 数据集中八个时长为一分钟的全景视频、32 位被试观看八个视频的视觉行为数据和实时连续情绪标注数据划分为五种时长（5s、10s、20s、30s、60s）的片段。表5.2中列出了不同时长片段下四种情绪标注类型的样本量以及样本总数量（片段数 × 视频数 × 被试数）。在所有时长的样本中，标签为

表 5.2 不同时长片段下四种情绪类型的样本数和样本总量

片段时长（秒）	样本大小				
	HAHV	**HALV**	**LALV**	**LAHV**	**总体（片段数 × 视频数 × 被试数）**
5	1073	826	419	754	12 × 8 × 32
10	487	43	236	381	6 × 8 × 32
20	224	225	124	195	3 × 8 × 32
30	145	146	92	129	2 × 8 × 32
60	70	74	47	65	1 × 8 × 32

“LVLA”情绪类型的样本数量最少。3.5.1小节已经验证了实时连续情绪标注数据的可信度和有效性，连续标注数据和 SAM 标注结果、原始数据集[63] 的情绪标签一致。本节计算每个片段中实时连续情绪标注序列唤醒值与效价值的平均值和中位数，以及头部运动、眼部运动的俯仰角与偏航角的平均数和标准差。W 检验表明这些片段中上述数据序列符合正态分布（$p > 0.05$），然后通过计算被试的头部运动与眼部运动数据和实时连续的唤醒-效价标注数据之间的皮尔逊积矩相关系数[63]，分析不同时长片段中二者之间的统计学相关性。

对于头部运动数据，每个时长类别的相关系数和对应的 p 值呈现在表5.3中，实验结果表明 5s、10s 和 20s 时长片段中被试的俯仰角均值与唤醒维度标注值的中位数之间存在中度显著正相关（$0.3 < corr < 0.6$，$p < 0.01$）；5s、10s 和 20s 时长片段的偏航角标准差与效价维度标注值的中位数之间存在中度显著负相关（$-0.6 < corr < -0.3$，$p < 0.01$）。

对于眼部运动数据，每个时长类别的相关系数和对应的 p 值呈现在表5.4中，实验结果表明 5s 和 10s 时长片段中被试的俯仰角均值与唤醒维度标注值的中位数之间存

表 5.3　头部运动和实时连续情绪标注之间的相关性及显著性表

片段时长（秒）	HM 数据	效价（均值）		效价（中位数）		唤醒（均值）		唤醒（中位数）	
		Corr	**p**	**Corr**	**p**	**Corr**	**p**	**Corr**	**p**
5	俯仰角（均值）	-0.147	0.154	**0.442**	**0.000**	-0.207	0.043	**0.458**	**0.000**
	偏航角（均值）	**0.306**	**0.002**	**0.317**	**0.002**	**0.264**	**0.009**	**0.264**	**0.009**
	俯仰角（标准差）	**0.304**	**0.003**	-0.249	0.015	0.243	0.017	-0.190	0.064
	偏航角（标准差）	0.051	0.620	**-0.315**	**0.002**	0.033	0.750	**-0.265**	**0.009**
10	俯仰角（均值）	-0.157	0.288	**0.483**	**0.001**	-0.229	0.117	**0.516**	**0.000**
	偏航角（均值）	0.318	0.027	0.327	0.023	0.299	0.039	0.242	0.098
	俯仰角（标准差）	0.335	0.020	-0.237	0.105	0.272	0.062	-0.112	0.448
	偏航角（标准差）	0.038	0.799	**-0.441**	**0.002**	0.002	0.989	-0.325	0.024
20	俯仰角（均值）	-0.167	0.436	**0.522**	**0.009**	-0.266	0.209	**0.585**	**0.003**
	偏航角（均值）	0.343	0.100	0.374	0.071	0.321	0.126	0.316	0.132
	俯仰角（标准差）	0.272	0.199	-0.263	0.214	0.135	0.530	-0.145	0.500
	偏航角（标准差）	0.085	0.692	**-0.532**	**0.007**	0.020	0.927	-0.417	0.042
30	俯仰角（均值）	-0.188	0.486	0.514	0.050	-0.287	0.282	0.560	0.024
	偏航角（均值）	0.381	0.145	0.363	0.167	0.350	0.184	0.333	0.208
	俯仰角（标准差）	0.370	0.158	-0.237	0.377	0.283	0.288	-0.136	0.615
	偏航角（标准差）	0.173	0.522	-0.547	0.028	0.149	0.581	-0.380	0.146
60	俯仰角（均值）	-0.194	0.645	0.517	0.189	-0.161	0.703	0.509	0.198
	偏航角（均值）	0.532	0.174	0.363	0.377	0.635	0.091	0.203	0.630
	俯仰角（标准差）	0.304	0.464	-0.402	0.323	0.320	0.440	-0.416	0.306
	偏航角（标准差）	0.227	0.588	-0.673	0.067	0.183	0.664	-0.508	0.199

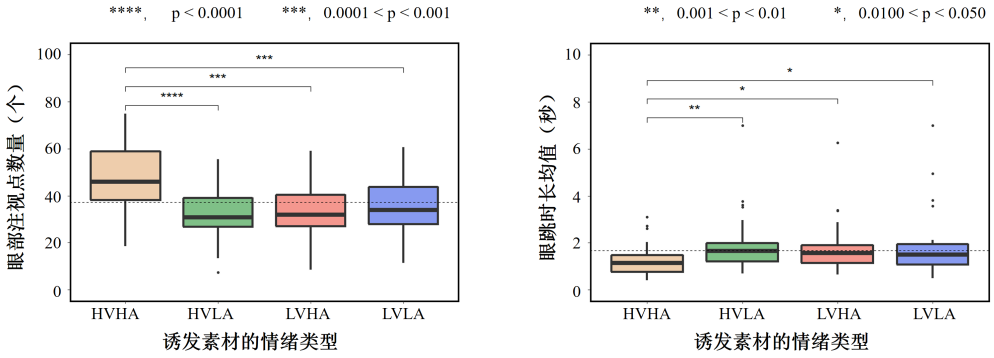
在中度显著负相关（$-0.6 < corr < -0.3$，$p < 0.01$）；5s 时长片段的偏航角标准差与效价维度标注值的中位数之间存在低度显著负相关（$-0.3 < corr < -0.1$，$p < 0.01$），10s 时长片段存在中度显著负相关（$-0.6 < corr < -0.3$，$p < 0.01$）。

对于注视和眼跳两个常见的眼部运动特征，本节计算了每位被试观看每个全景视频时眼部注视点数量，注视时长的总长度、标准差、均值；眼跳时长的均值与标准差。实验结果表明被试在观看不同情绪类型视频的过程中注视点数量均值之间存在显著性差异（$p < 0.001$），情绪类型为“HVHA”的视频中被试注视点数量高于其他情绪

表 5.4　眼部运动和实时连续情绪标注之间的相关性及显著性表

片段时长（秒）	EM 数据	效价（均值）		效价（中位数）		唤醒（均值）		唤醒（中位数）	
		Corr	**p**	**Corr**	**p**	**Corr**	**p**	**Corr**	**p**
5	俯仰角（均值）	0.068	0.512	**-0.321**	**0.001**	0.136	0.185	**-0.354**	**0.000**
	偏航角（均值）	**0.299**	**0.003**	**0.275**	**0.007**	0.257	0.011	0.223	0.029
	俯仰角（标准差）	**0.286**	**0.005**	-0.248	0.015	0.223	0.029	-0.166	0.106
	偏航角（标准差）	0.144	0.161	**-0.284**	**0.005**	0.123	0.234	-0.200	0.051
10	俯仰角（均值）	0.073	0.624	-0.360	0.012	0.162	0.273	**-0.421**	**0.003**
	偏航角（均值）	0.314	0.030	0.288	0.047	0.293	0.043	0.200	0.174
	俯仰角（标准差）	**0.395**	**0.005**	-0.307	0.034	0.339	0.018	-0.178	0.226
	偏航角（标准差）	0.124	0.400	**-0.418**	**0.003**	0.083	0.575	-0.276	0.058
20	俯仰角（均值）	0.075	0.728	-0.390	0.060	0.191	0.370	-0.472	0.020
	偏航角（均值）	0.339	0.105	0.332	0.113	0.316	0.133	0.271	0.201
	俯仰角（标准差）	0.305	0.147	-0.350	0.094	0.170	0.426	-0.194	0.363
	偏航角（标准差）	0.175	0.414	-0.429	0.036	0.108	0.615	-0.297	0.159
30	俯仰角（均值）	0.091	0.738	-0.373	0.154	0.211	0.432	-0.443	0.085
	偏航角（均值）	0.383	0.144	0.318	0.230	0.351	0.182	0.282	0.290
	俯仰角（标准差）	0.374	0.154	-0.354	0.178	0.268	0.315	-0.218	0.417
	偏航角（标准差）	0.239	0.373	-0.436	0.091	0.216	0.423	-0.248	0.354
60	俯仰角（均值）	0.089	0.833	-0.382	0.350	0.093	0.826	-0.394	0.335
	偏航角（均值）	0.555	0.153	0.315	0.448	0.656	0.077	0.146	0.730
	俯仰角（标准差）	0.381	0.351	-0.537	0.170	0.345	0.403	-0.525	0.181
	偏航角（标准差）	0.345	0.403	-0.586	0.127	0.282	0.499	-0.399	0.327

类型的视频，结果呈现图5.5(a)；被试在观看不同情绪类型视频的过程中眼跳时长均值之间存在显著性差异（$p < 0.05$），“HVHA”类型的视频中眼跳时长低于其他类型的视频，结果呈现在图5.5(b)中。



(a) 注视点数量均值　　(b) 眼跳时长的均值

图 5.5　不同情绪类型的视频之间眼部运动特征直方图

5.5.3　视觉交互行为与情绪的相关性讨论

首先，针对唤醒维度的实时连续情绪报告，如图5.6所示，5s、10s 和 20s 时长片段的头部运动俯仰角均值与唤醒度中位数之间存在正相关，这表明被试在报告高唤醒值时通常会抬头，在报告低唤醒值时会相对低头；这与 Li 等人[63] 的实验结果一致，同时 Lhommet 等人[227] 的研究也发现人们在感到担心或是惊讶时（对应高唤醒值）往往会向后仰头。但是对于眼部运动数据，5s 和 10s 片段的俯仰角均值与唤醒度中值呈负相关。由于用户在报告高唤醒值时通常抬起头，而全景视频的核心内容往往呈现在赤道附近区域[51,228]，因此用户在抬头的同时眼睛会向下看，在低头时会向上看。

其次，对于效价维度的实时连续情绪报告，如图5.7所示，实验结果指出 5s、10s 和 20s 时长片段的头部运动偏航角标准差与效价度中位数之间存在负相关，这表明当用户在报告较低的效价值时，通常伴随有较为明显的头部左右旋转；这一结果与 Won 等人[177] 报告的头部偏航运动数量与焦虑值之间的显著相关性一致，但与 Li 等人[63] 报告的结果相反。可能的原因是 Li 等人在研究中采用的诱发视频时长均大于两分钟，长时间观看会带来疲惫感和眩晕感，因此用户在观看过程中仅关注直接呈现在眼前的内容，基本没有探索行为；此外，Li 等人实验中用户的情绪报告发生在视频观看完成之后，这也会对情绪报告的精确性产生影响。本研究中诱发视频时长均为一分钟，被试在虚拟空间中探索内容的同时实时报告情绪状态，因此实验中划分的短时长片段里（30 秒以内），被试在转动头部寻找目标点时报告了较低的效价值。同样地，5s 和 10s 时长片段眼部运动数据的偏航角标准差与效价度中位数之间呈负相关。

第三，Tang 等人[38] 指出消极的情绪状态对注视和眼跳特征具有显著影响，更大、

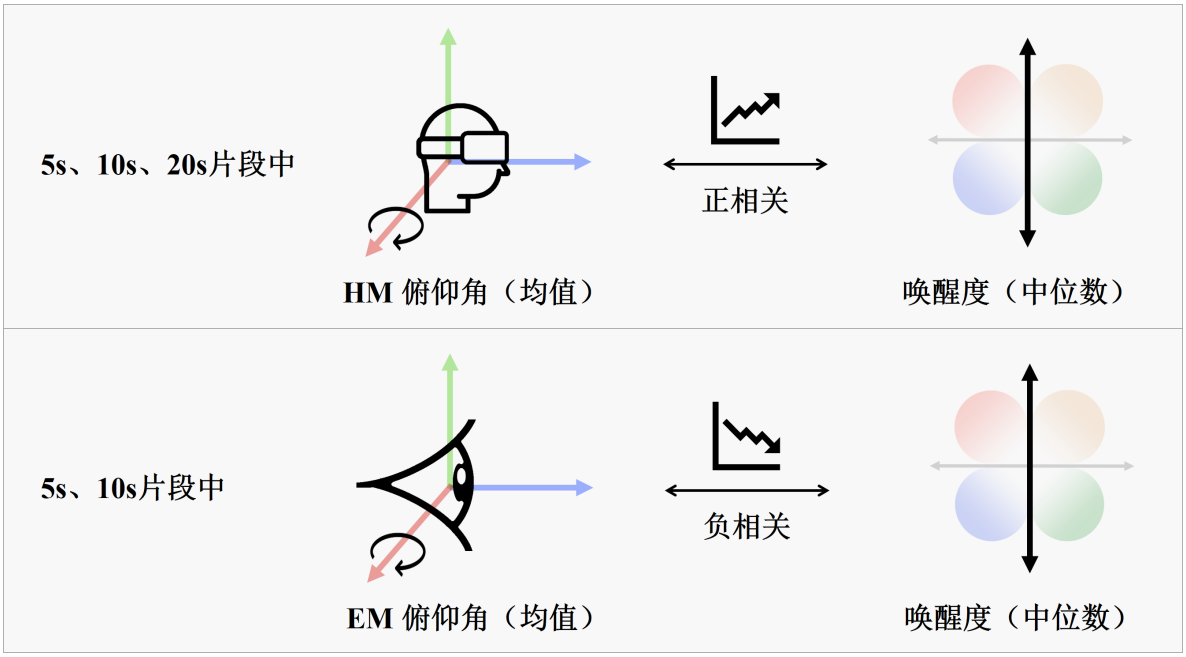


图 5.6　头部运动与眼部运动和情绪唤醒维度之间的相关性

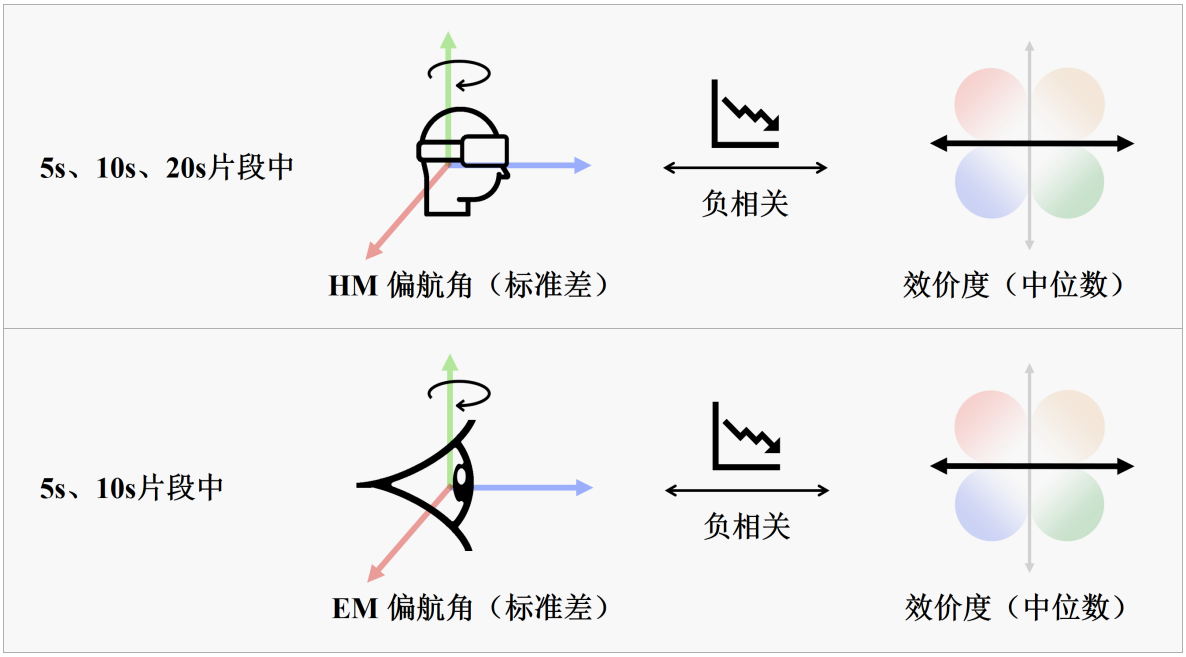


图 5.7　头部运动与眼部运动和情绪效价维度之间的相关性

更长和更快的眼跳容易产生视觉扰动和回避行为。本研究中情绪类型为“LV（低效价值）”的视频眼部特征结果与 Tang 等人的结论一致；相比于“HVHA”情绪类型的视频，用户在观看“LVHA”与“LVLA”类型视频时表现出更少的注视行为和更长的眼跳时长。然而，“HVLA”相比“HVHA”具有更少的注视行为和更长的眼跳时长，这是由于诱发素材中两个情绪类型为“HVLA”的视频内容分别是海边日出（$TpI = 0.36$）

和雪山场景（$TpI = 0.97$），时间复杂度相对较低，因此用户在体验过程中也更多地在探索视频内容。

5.6 本章小结

本章首先提出了虚拟环境中视觉交互行为的主要研究对象是用户的头部运动与眼部运动，介绍了二者在真实环境、虚拟空间、等距投影三种情境下的转换方式，从头部运动与眼部运动原始数据中提取俯仰角和偏航角的计算方法、注视和眼跳两种常见的眼动行为特征获取方法、以及头部运动的扫描路径。研究采用二维显著图方法对用户在虚拟体验中的视觉注意进行建模，探讨了虚拟环境中用户的四种视觉行为特征：用户之间的头部运动与眼部运动一致性、用户头部运动与眼部运动之间的相关性、用户视觉行为的赤道及前方偏向、诱发素材内容对用户视觉行为的影响。视觉交互行为特征实验结果显示：（1）用户在观看全景视频时的头部运动和眼部运动均具有很高的一致性；（2）尽管用户的视觉显著区域会受到观看内容的影响，但仍然能够观察到明显的赤道和前方偏向；（3）对于头部运动的俯仰角，被试 90% 以上的时间集中在-30° 到 30° 区域之间，对于偏航角，被试观看-150° 到 150° 区域的时间不足 10%。本章还提出一种虚拟环境中细粒度层级的用户视觉交互行为与实时连续情绪的相关性评估方法，实验结果表明：（1）头部运动偏航角的标准差与效价值呈负相关，而头部运动俯仰角与唤醒值呈正相关；（2）眼部运动偏航角的标准差与效价值呈负相关，而眼部运动的俯仰角与唤醒值呈负相关；（3）对于“HVHA”情绪类型的视频，用户的眼部注视数量多，而眼跳时长短。

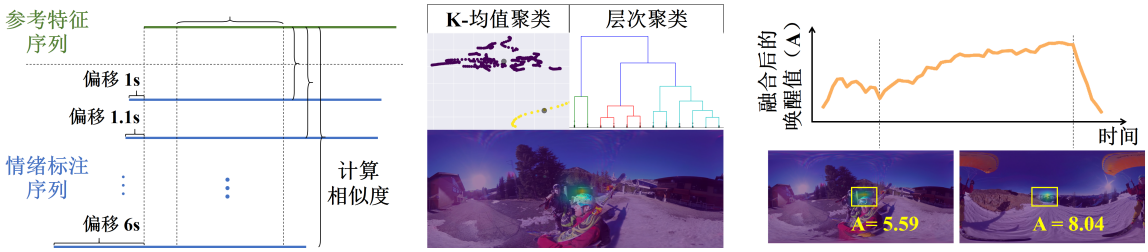
本研究为理解用户在虚拟交互体验中的头部运动和眼部运动行为特征及其与实时连续情绪标注之间的关系提供了基础，能够实现以一种低成本方法，在细粒度层级上提高沉浸式虚拟交互过程中的用户体验质量。

第 6 章　基于视觉交互行为的情绪识别

6.1　引言

情感计算领域的一个关键步骤是定义精确的情绪标签，用于理解情绪状态、训练和测试情绪识别模型。在情绪监测与评估过程中，实时连续的情绪报告相比于离散情绪标签能够反映更精细的时变信息。为了更好地理解不同用户针对同一诱发素材的情绪状态，构建用户独立的情绪识别模型，需要将多个用户实时连续的情绪标注序列融合成为一组情绪标签。值得注意的是，人从观察感知诱发素材到给出情绪报告之间存在一定的反应延迟，延迟时长受到个体性别、年龄、注意力等因素影响，在个体之间存在差异[20]。特别是在虚拟环境中，用户能够自由转动头部选择视口区域，因此每个时刻的情绪标注是根据个体视口相关的体验内容产生的[170]，传统的求取所有用户情绪序列均值方法[229]没有考虑给定时刻情绪数据对应的场景信息。

为此，本章构建了不同视觉交互行为下用户的情绪识别研究方法，如图6.1所示，以确保虚拟环境中用户实时连续的标注数据能够建立用户独立的情绪 Ground-Truth 标签。主要研究内容分为如下三个方面：



(a) 实时连续情绪信号时间对齐　(b) 细粒度层级的视口聚类　(c) 细粒度层级的情绪标注数据融合

图 6.1　基于视觉交互行为的情绪识别流程图

（1）提出一种实时连续情绪信号的时间对齐方法，从诱发素材中提取与情绪相关性最高的主要特征作为参考特征序列，通过滑动窗口算法和动态时间规整获取并衡量参考特征序列与不同偏移下的情绪标注序列之间的相似度，相似度最高的标注序列对应的偏移量即为该序列的反应延迟时长。

（2）提出一种用户视口相关的实时连续情绪融合方法，将诱发素材划分为若干等时长片段，对每个片段中用户的头部运动行为数据进行 K-均值聚类或层次聚类，获

取每个片段视口聚类的用户簇，对聚类簇中用户时间对齐的情绪数据进行单元级和片段级融合。

（3）研究分析基于视觉交互行为的情绪融合方法的有效性及合理性，测试并评估CEAP-360VR 数据集中 32 位被试观看八个不同情绪类型视频的情绪融合结果，探索融合后情绪标签的准确性并在细粒度层级上对情绪标签的峰谷、变化趋势及其与对应场景之间的相关联系进行时序性分析。

实验结果表明，用户在全景视频体验中的头部运动具有较高的一致性，80% 视频片段的聚类结果中包含了 18 位以上（$N = 32$）的被试；融合后的实时连续情绪标签能够精准地分类或预测诱发素材的原始标签及离散标签，并能够提供情绪状态的峰值、波谷、变化趋势等时序细节信息，及其与对应场景之间的相关联系。以下将会详细描述基于视觉交互行为的情绪识别方法和评估实验。

6.2 相关工作

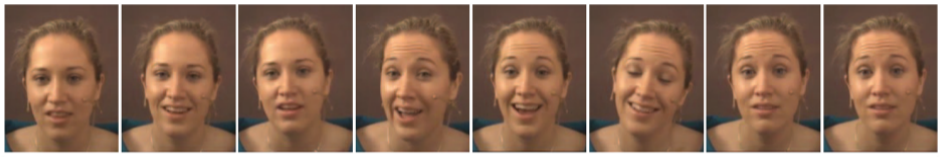
为了从多个用户的实时连续标注序列中获取精确的情绪标签，首先需要计算情绪标注内容与标注数据之间的时间延迟，以及标注员之间的反应延迟差异。Huang 等人[19] 通过去除标注序列的前 N 帧与特征序列的后 N 帧实现二者时间对齐，用于训练回归模型生成情绪预测值，并采用平滑滤波器重新对齐预测标签与基准标签。Nicolaou 等人[230] 提出一种基于相关性的特征选择方法，利用标签与特征之间的相关系数构建延时概率分布用于特征选择，并验证了该方法的有效性和鲁棒性。Mariooryad 等人[229] 通过最大化表情行为信号与实时连续情绪标注之间的互信息（Mutual Information），获取每个标注员的延迟时长，并对标注数据进行平移对齐弥补标注延迟造成的影响，实验表明该方法相比基线分类器提高了 7% 的准确率。为了融合来自多个标注员情绪数据形成一组情绪标签，Mariooryad 等人[229] 在研究中计算了所有标注员经时间对齐处理后标注数据的平均值。Sakoe 等人[231] 于 1978 年提出的动态时间规整（Dynamic Time Warping, DTW）是另一个常用的时间对齐方法，通过动态规划和时间扭曲最大化特征序列与标注序列之间的相似度，尽可能实现两个序列的时间对齐。Nicolaou 等人[232] 基于典型相关性分析（Canonical Correlation Analysis，CCA）提出了一种新的回归模型：相关空间回归（Correlated-Spaces Regression, CSR），该方法运用各维度之间的相关性并同时实现了特征序列的有监督降维和多模态融合，提高了融合精度。这些研究均是在非沉浸式环境中进行。

在虚拟环境中，用户能够通过头部运动自由地选择观看视口，其情绪是基于视口内容产生的。为了获取虚拟环境中多个用户针对同一诱发素材精确的情绪标签，需要理解用户如何探索诱发素材以及如何与之交互，并开发新的方法量化、分析、识别虚拟环境中用户视口相关的情绪标签。Marmitt 等人[233] 指出头部运动和眼部运动数据常用于虚拟环境中观看行为的分析研究。Wu 等人[234] 构建了 48 位被试观看多种类型全景视频时的头部运动数据集；Xu 等人[201] 分析了 58 名被试观看 76 个全景视频时的头部运动数据，研究结果均指出用户在观看全景视频时具有相似的视觉交互行为模式。Rossi 等人[235] 提出一种基于图的聚类方法，通过计算大圆距离衡量不同用户之间的观看相似性，用于识别并构建同一时刻观看视频同一区域的用户簇。Nasrabadi 等人[236] 对用户观看区域进行聚类分析，并提出一种基于视口的预测方法。这些工作推动了虚拟环境中用户视觉行为建模和视觉注意预测的研究，但现有的研究均未考虑用户在视口变化过程中的情绪标注问题，没有解决不同视觉交互行为下用户情绪识别的相关问题。

6.3 情绪信号时间对齐

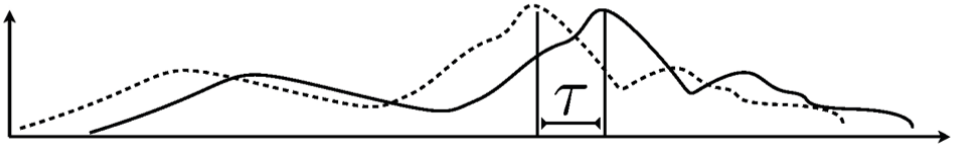
实时连续情绪识别的一个关键问题在于计算情绪诱发内容与情绪标注信号之间的时间延迟。在感知评估过程中，人在接收外界的诱发刺激后，需要经过大脑加工处理转换成机体内在的情感活动，进而完成实时连续的情绪标注。因此从诱发事件发生到标注者完成内容的情绪标注之间存在时间延迟，并且这一反应延迟的时长受到个体性别、年龄、注意力等因素影响，在个体之间存在差异[20]，如图6.2所示。一项实证分析[229] 表明，考虑标注者的个体因素、标注维度数量及视觉行为方式，人的反应延迟时长通常在 1 至 6 秒之间。Mariooryad 等人[229] 提出一种用户独立的反应延迟计算方法，通过计算诱发素材的情绪数据与用户标注数据之间的交互信息量获取标注者相关的延迟时长；动态时间规整（DTW）算法[231] 通过动态规划的方法计算不同长度时间序列的距离，常用于衡量时序性数据之间的相似性。基于此，本节提出一种实时连续情绪信号的时间对齐方法，具体步骤如下：

（1）获取诱发素材的参考特征序列，Soleymani 等人[25] 提出用户的生理信号可用于对齐情绪标注序列，但每个人具有不同的生理水平和反应时长，因此生理信号无法作为所有用户标注数据的对齐标准。本研究首先从诱发素材中选择并提取与情绪相关的主要特征，选取的原则是基于诱发素材特点和先验经验，例如视频类内容考虑视觉



(a) 诱发素材中演讲者的面部表情

(b) 诱发素材中演讲者的音频信号



(c) 实线表示情绪标注序列，虚线表示诱发素材的情绪信号，τ 为二者之间的延迟时长

图 6.2　情绪标注反应延迟说明[25]

要素（颜色、纹理等）、音乐类内容考虑声音信息（频率、响度等）、对话类内容考虑讲话者的面部表情和肢体动作等；然后计算选取的主要特征序列与情绪标注序列之间的相关性，选择与情绪相关性最高的主要特征作为参考特征序列（Reference Feature，RF）。由于虚拟空间中用户能够自由地转动头部选择视口位置，如果将视觉要素作为主要特征，首先应选取用户头部运动没有较大变化的时间区间，并从对应的视口区域中提取视觉信息作为特征序列；例如，用户在虚拟体验的初始几秒内头部运动通常较为稳定，相似度高，可以作为备选时间段。假设 RF_j 是从第 $j \in [1, J]$ 个诱发素材中提取的参考特征，$RF_j = [RF_j^1, RF_j^2, \cdots, RF_j^N]$，其中 J 表示诱发素材的总数量，N 表示参考特征序列 RF_j 的总帧数。

（2）计算情绪标注序列的延迟时长，假设 P_{ij} 是用户 $i \in [1, I]$ 在参考特征序列 RF_j 对应的时间区间内（诱发素材 j）的情绪标注数据（唤醒或效价维度），$P_{ij} = [P_{ij}^1, P_{ij}^2, \cdots, P_{ij}^M]$；$D_{ij}$ 表示用户 $i \in [1, I]$ 在体验诱发素材 $j \in [1, J]$ 时的反应时长；fps_j 为第 j 个诱发素材的采样频率，I 和 M 分别表示用户的总数量和诱发素材 i 的总帧数。首先采用滑动窗口（Sliding Window）算法获取偏移后的实时连续情绪标注序列，滑动窗口大小设置为与参考特征序列时长一致，滑动步长根据情绪标注数据的采样频率设置为其倒数，研究[229] 表明用户的反应延迟时长为 1-6 秒，因此滑动总步长的区间为 $[1, 6]$，偏移后的标注序列记为 S_P_{ij}。然后，采用 DTW 方法衡量参考特

征序列 RF_j 与偏移后的标注序列 S_P_{ij} 之间的距离 Dis_τ，距离越短表明两个序列之间的相似度越高；相似度最高的标注序列对应的总滑动步长 τ 即为该标注序列的反应延迟时间 D_{ij}，计算标注延迟时长的伪代码见算法1。

Algorithm 1 标注延迟时长计算

Input: 唤醒-效价情绪标注序列 $P \in R^{I\times J}, P_{ij} = [P_{ij}^1, P_{ij}^2, \cdots, P_{ij}^M]$;

参考特征序列 $RF \in R^{1\times j}, RF_j = [RF_j^1, RF_j^2, \cdots, RF_j^N]$

Output: 每位被试针对每个诱发素材的标注延迟时长 $D \in R^{i\times j}$

for j = 1 to I **do**
  for i = 1 to J **do**
    for τ = 1 to 6, 步长为 0.1 **do**
      $S\_P_{ij} = [P_{ij}^{1+\tau*fps_j}, P_{ij}^{2+\tau*fps_j}, \cdots, P_{ij}^{N+\tau*fps_j}]$
      $Dis_\tau = DTW(S\_P_{ij}, RF_j)$
    end for
    $D_{ij} = argmin(Dis)$
  end for
end for

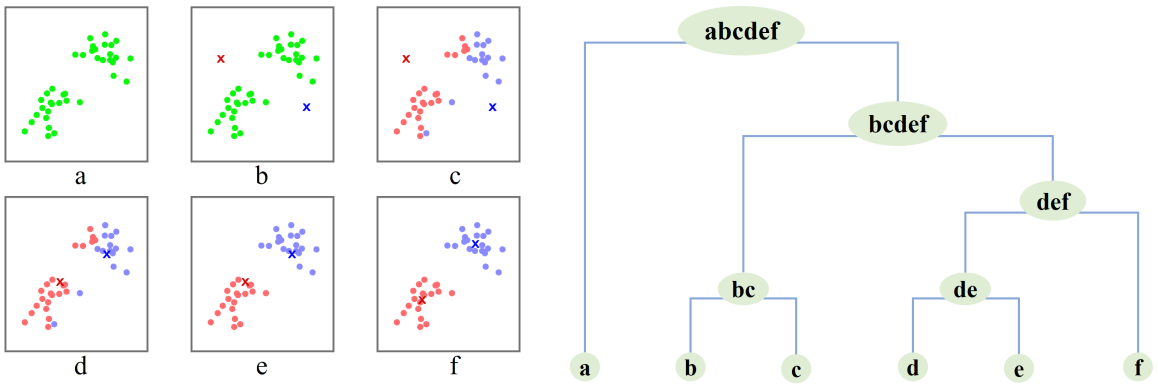
（3）记录时间对齐后的情绪标注序列，根据上述步骤中的反应延迟时间 D_{ij} 向后偏移情绪标注序列，获取用户 $j \in [1, J]$ 针对诱发素材 $i \in [1, I]$ 时间对齐后的实时连续情绪标注序列。

6.4 视口相关的实时连续情绪融合

在虚拟环境中，用户的情绪标注是由其选择观看的视口内容驱动的。为了融合不同用户时间对齐后的实时连续情绪标注结果，构建有效的情绪 Ground-Truth 标签，本节对所有用户的视觉交互行为模式进行片段级视口聚类，获取具有相似观看行为的用户簇；并对每个片段聚类结果最大簇中的情绪标注数据进行单元级和片段级融合。

6.4.1 片段层级视口聚类

虚拟环境中的用户视觉注意研究[212]表明，用户在虚拟体验中的视觉行为模式具有较高的一致性（见5.3.2小节）。本小节采用片段级的聚类方法获取具有相同观看行为的用户。首先将诱发素材 j 划分为时长为 t 的若干个片段序列，$Seg_j = [Seg_j^1, Seg_j^2, \cdots, Seg_j^S]$，其中 S 表示诱发素材 j 的片段总数量；将每个片段内所有用户在体验过程中的头部运动采样点进行聚类，形成 C 个簇（$C \geq 1$），确保每个片段中数据点最多的簇内有超过 80% 的头部运动采样点。获取每个簇内采样点对应的用户编号，同时记录每个簇的质心位置作为该片段内用户视口聚类后的中心点位置。考虑以下两种常见的聚类方法：



(a) K-均值聚类流程图　　(b) 层次聚类流程图

图 6.3　两种常见的聚类方法

（1）K-均值聚类（K-Means Clustering）[237]，该方法首先需要指定聚类后得到的用户簇数量，即 K 值（$K \geq 2$），随机选取 K 个样本点作为 K 个簇的初始中心，将每个样本点归为距离其最近的中心点所在簇，所有样本点归类完毕后重新计算 K 个簇的中心，重复上述过程至每个簇内的样本点不再变动（或是满足某种条件规则），这一过程呈现在图6.3(a)中。K-均值聚类方法简单快速，适合处理大规模密集型数据，但事先指定的簇数量和中心点会对聚类结果产生影响，且不适用于不规则簇。

（2）层次聚类（Hierarchical Clustering）[238]，不同于 K-均值聚类，层次聚类不需要提前指定集群数量，该方法将每个样本点视为一个簇，计算各个簇之间的距离，最近的两个簇合成为一个新簇，重复上述过程至只剩最后一个簇或是满足某种条件规则，如图6.3(b)所示。层次聚类方法作为一种自下而上的聚类方法，适用于任意形状的聚

类，在聚类过程中也能够动态调整收敛距离参数；但该方法的时间复杂度大，且每一次聚类是基于上一次聚类结果进行的，具有不可逆性[239]。

根据数据类型选择适当的聚类方法，获取诱发素材每个片段中聚类簇的用户数量和用户编号，并提取这些用户时间对齐的标注序列（见6.3节）。

6.4.2 情绪标注信号融合

本节将通过以下两个步骤完成虚拟环境中视口相关的实时连续情绪标注信号融合：（1）单元级（每个采样点）融合，该步骤在细粒度层级对实时连续标注数据进行清洗，去除序列中的噪声和离群值后进行融合；（2）片段级融合，对标注序列进行融合，融合算法的伪代码见算法2。

Algorithm 2 视口相关的标注融合

Input: 唤醒-效价情绪标注序列 $P_{is} = [P_{is}^1, P_{is}^2, \cdots, P_{is}^N], P \in R^{I \times S}$

Output: 融合后的唤醒-效价情绪标注序列，$F \in R^{1 \times S}$

for s = 1 to S **do**
  for n = 1 to N **do**
    for i = 1 to I **do**
      P_{is}^n 的 D^n，根据 **公式 6.1**
    end for
    $X_s \leftarrow$ 删除 P_s 中的 P_{is}^n，当 $d_{lm}^n > T$
    $f_n \leftarrow$ 融合 X_s，根据 **公式 6.2**
  end for
  $F_s = \sum_{n=1}^{N} \frac{H_n}{\sum_{p=1}^{N} H_p} f_n$
end for

假设 P_{is} 是用户 $i \in [1, I]$ 在诱发素材第 $s \in [1, S]$ 个片段体验过程中唤醒或效价维度的连续情绪标注数据（已完成时间对齐处理），$P_{is} = [P_{is}^1, P_{is}^2, \cdots, P_{is}^N]$，其中 I 表示进行融合的用户总数量、S 表示诱发素材的片段总数量、N 表示标注序列片段 s 的采样点总数量。首先采用贝叶斯融合方法（Bayesian Fusion）[240,241]，对多个用户的标注数据在每个采样点进行单元级融合，置信测度矩阵记为 D^s，其中第 $n \in [1, N]$ 个采

样点的 $d_{lm}^{n} \in D^{n}$ 计算方式如下：

$$d_{lm}^{n} = erf(\frac{x_l - x_m}{\sqrt{2}\sigma_l}), \qquad d_{ml}^{n} = erf(\frac{x_m - x_l}{\sqrt{2}\sigma_m}) \tag{6.1}$$

上式中，x_m 与 x_l 分别表示用户 m 与用户 l 的标注数据，σ_m 与 σ_m 分别表示一个片段中用户 m 与用户 l 标注数据的标准差；$erf(\theta) = \frac{2}{\pi}\int_{\theta}^{0} e^{-u^2}$ 是误差函数。通过把 d_{lm} 的阈值设置为 $T = 0.2$ 去除标注数据在第 n 个采样点的离群值，去除离群值后的标注序列记为 $X_s = [x_1, x_2, \ldots, x_Q]$。第 n 个采样点的标注融合结果计算如下：

$$f_n = \sum_{q=1}^{Q}(1 - \frac{\sum D_q^s}{\sum D^s}) \cdot x_q \tag{6.2}$$

上式中，D_q^s 表示置信测度矩阵 D^s 的第 q 列，每个片段中所有采样点 $n \in [1, N]$ 在单元级的融合结果记为 $f = [f_1, f_2, \ldots, f_N]$。诱发素材片段 s 中标注数据的片段级融合采用 f 的加权平均数计算如下：

$$F_s = \sum_{n=1}^{N} \frac{H_n}{\sum_{p=1}^{N} H_p} f_n \tag{6.3}$$

上式中，H_n 表示第 n 个采样点参与聚类的视口信息数量。分别计算诱发素材中每个片段 $s \in [1, S]$ 中融合后的标注数据，获取该诱发素材视口相关的实时连续情绪标注融合序列。由于唤醒和效价是两个正交独立变量，因此这两个维度的标注数据需要独立计算。

6.5 实验结果与讨论

为了验证上述基于视觉交互行为的情绪识别方法的有效性和合理性，本节测试并评估了这些方法在 CEAP-360VR 数据集中的情绪融合结果，对 32 位被试观看八个不同情绪类型视频过程中的实时连续情绪标注序列进行细粒度层级视口相关的融合。本章实验结果主要从实时连续的情绪数据时间对齐、视口相关的情绪数据融合以及实验结果讨论三个方面逐一进行详述。

6.5.1 情绪数据时间对齐结果

首先从诱发素材中选取参考特征。用户在观看全景视频时能够通过头部运动自由选择观看区域，CEAP-360VR 数据集中的八个全景视频内容涉及飞行体验、动作短片、风景欣赏等多个类别，音频信息包括对话、纯音乐、独白等多种类型（见表3.2），为了消除音视频内容对参考特征选取的影响，本实验考虑被试观看行为相似度最高的时间段，并提取视频在该时间段内每一帧图像中的颜色（Color）、纹理（Texture）和边缘（Edge）三种常见视觉要素作为特征序列，研究[242–245] 表明上述三个视觉特征能够影响人的情绪状态，Xu 等人[171] 指出这些特征同样适用于全景视频观看中用户情绪研究。在 CEAP-360VR 数据集中，所有被试的观看方向初始化在视频中心点（见3.4.2

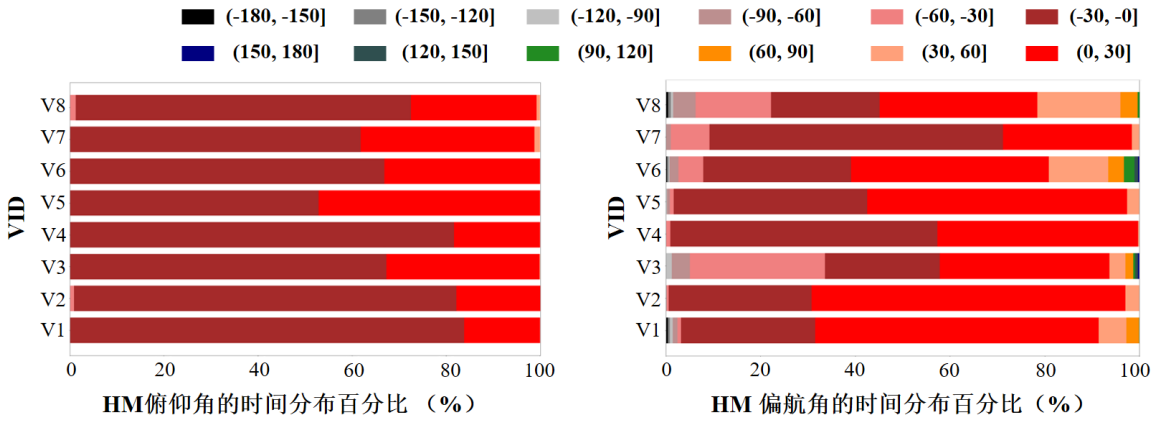


图 6.4 诱发素材前六秒中头部运动的俯仰角和偏航角时间分布情况

（4）），实验选取每个视频的前六秒内容，计算所有被试头部运动在此区间内的偏航角和俯仰角分布情况，结果呈现在图6.4中：在八个视频播放的前六秒，所有被试头部运动俯仰角在 98% 以上的时间内位于-30° 到 30° 之间（图6.4左），偏航角在 90% 以上的时间内位于-60° 到 60° 之间（图6.4右）。因此选择每个视频前六秒内每一帧的特定区域（偏航角 $[-60°, 60°]$，俯仰角 $[-30°, 30°]$），并从中提取图像的颜色、纹理和边缘信息。采用 Stricker 等人[246] 提出的颜色矩（Color Moment）方法计算每帧图像的一阶、二阶、三阶颜色矩，该方法可以有效地表示图像中颜色分布，$CF_i = [CF_i^1, CF_i^2, \cdots, CF_i^N]$ 是从视频 $i \in [1, I]$ 中提取的颜色特征，其中 I 表示视频总数量、N 表示视频 i 前六秒的总帧数；采用灰度共生矩阵（Gray-Level Co-occurrence Matrix，GLCM）[247] 方法提取纹理特征 TF_i，该方法通过研究图像灰度的空间相关特性来描述纹理；采用 Canny 算子（Canny Operator）[248] 提取图像的边缘特征 EF_i。W 检验表明诱发素材的上述三

个视觉特征序列均不符合正态分布（$p < 0.05$）。

接下来，对所有被试的实时连续情绪标注序列重采样至与诱发素材帧率一致，假设 $P_{ij} = [P_{ij}^1, P_{ij}^2, \cdots, P_{ij}^N]$ 为被试 $j \in [1, J]$ 在全景视频 $i \in [1, I]$ 前六秒中每一帧的情绪标注数据序列（唤醒或效价维度），分别计算每个视频的三种特征序列 CF_i、TF_i、EF_i 和每个被试原始情绪标注序列 P_{ij} 的 Spearman 相关系数，结果呈现在表6.1中。八个全景视频的颜色特征 CF 和唤醒维度（ρ：$[0.226, 0.634]$）与效价维度（ρ：$[0.213, 0.618]$）的相关性最高，因此选择每个视频前六秒的颜色特征 CF_i 作为参考特征用于标注序列时间对齐处理。

表 6.1　参考特征序列与情绪标注序列的 Spearman 相关系数

	效价			唤醒		
	CF	**TF**	**EF**	**CF**	**TF**	**EF**
V1	**0.600**	0.549	0.502	**0.520**	0.473	0.437
V2	**0.618**	0.528	0.550	**0.634**	0.531	0.559
V3	**0.465**	0.430	0.092	**0.374**	0.350	0.112
V4	**0.400**	0.375	0.327	**0.464**	0.445	0.419
V5	**0.572**	0.522	0.206	**0.518**	0.461	0.171
V6	**0.213**	0.160	0.083	**0.226**	0.178	0.091
V7	**0.517**	0.467	0.476	**0.564**	0.506	0.549
V8	**0.577**	0.421	0.330	**0.583**	0.481	0.278

以标注序列 P_{ij} 的 $1-6s$ 为初始点，采用长度为 $6s$、步长为 $0.1s$（与标注设备 Joy-Con 的采样频率一致，见3.4.1.3（4））的滑动窗口，总移动步长记为 τ，获得移动 τs 后的情绪标注序列 S_P_{ij}；采用 DTW 方法计算 S_P_{ij} 与 CF_i 之间的距离，距离最短的 S_P_{ij} 对应的 τ 值即为该标注序列的延迟时长。结果表明八个视频中所有被试在效价维度上标注延迟时长（单位是秒）的均值范围是 $[1.5, 4.0]$（$M = 3.186, SD = 0.809$），在唤醒维度上标注延迟时长的均值范围是 $[2.0, 3.5]$（$M = 2.909, SD = 0.614$）；32 位被试针对所有全景视频在效价维度上标注延迟时长的均值范围是 $[2.0, 4.6]$（$M = 3.186, SD = 0.751$），在唤醒维度上标注延迟时长的均值范围是 $[1.6, 3.9]$（$M = 2.909, SD = 0.553$）。

6.5.2 视口相关情绪融合结果

本节对 CEAP-360VR 数据集中 32 位被试观看八个全景视频的头部运动进行聚类分析后，对每个聚类簇中用户的实时连续情绪标注序列进行融合，并展示了融合结果。

根据 Xie 等人[249] 的研究，实验将每个视频以及用户标注序列和行为数据划分为时长为 1 秒的片段，共 60 个；对每个片段内所有被试头部运动的采样点进行聚类分析。首先采用 K-Means 聚类方法检查用户在观看过程中头部运动数据的分布，计算八个视频在 K 的不同取值下（$K \in [2,8]$）最大用户簇（Majority Cluster）中被试数量分布情况，结果如图6.5所示。当 $K = 2$ 时，八个视频聚类后的最大簇中用户数量明显多于其他取值结果，所有视频中 80% 的片段中包含了至少 18 位被试；K 值越小，最大簇中所包含的用户数量越多。因此，对于每个诱发素材的所有片段，被试的头部运动数据聚类结果均存在一个明显的最大用户簇。

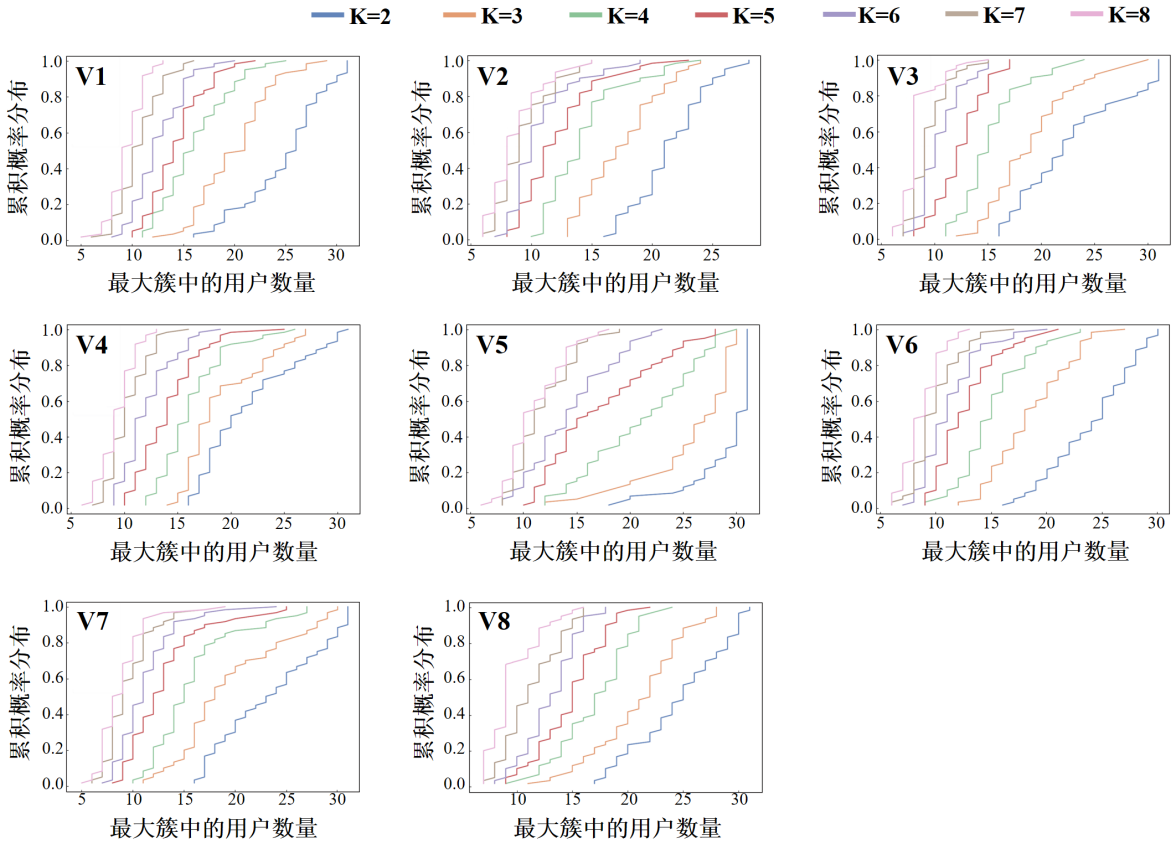
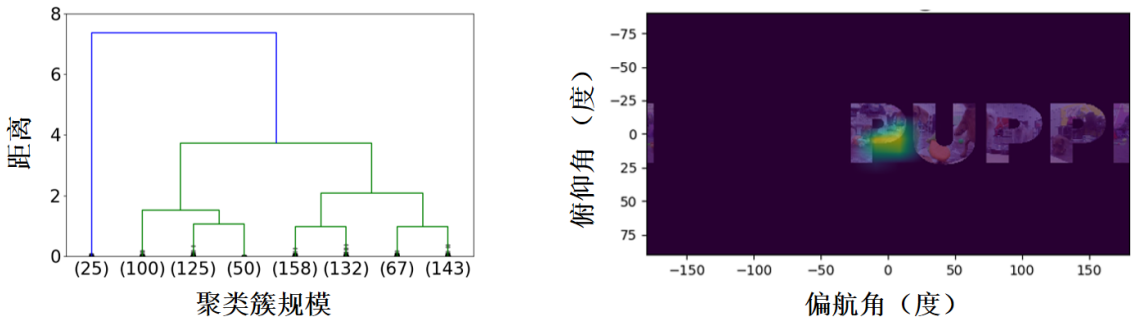


图 6.5 K-Means 聚类的最大簇中被试数量 CDF 分布情况

为了获取更精确的聚类结果用于不同用户的情绪标注序列融合，实验采用层次聚类方法，通过动态调整层次聚类的收敛距离，确保每个最大簇中包含 80% 以上的头部

运动采样点。图6.6(a)所示为视频 V1 中片段 1 的层次聚类树状图结果，图6.6(b)为该片段中所有被试头部运动数据点生成的显著图。八个全景视频中所有片段最大簇包含



(a) V1 中片段 1 的层次聚类树状图　　(b) V1 中片段 1 的 HM 显著图

图 6.6　层次聚类树状图和结果显著图示例

的被试数量占被试总数量的比例呈现在图6.7中，横轴为片段编号，范围是 $[1, 60]$，纵轴为最大簇中被试数量占比，范围是 $[0.4, 1.0]$；从图中可以看出，所有视频片段的最大簇中包含了 50% 以上的被试。八个视频所有片段的用户视口层次聚类分布情况如

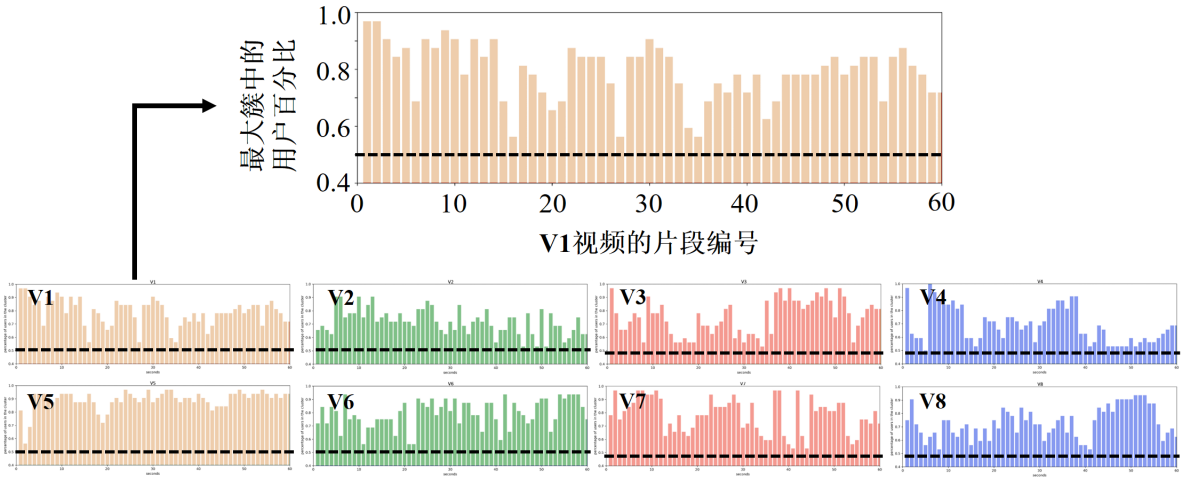


图 6.7　层次聚类的被试数量占比情况

图6.8所示，实验结果表明：（1）八个视频所有片段用户视口聚类的结果中包含了 16 位以上（$>$ 50%）的被试；（2）八个视频 80% 以上片段用户视口聚类的结果中包含了 18 位以上的被试；（3）视频 V5 所有片段中被试的头部运动一致性最高，其余七个视频所有片段最大簇中的用户数量较为相似。

完成层次聚类后，获取八个视频每个片段中最大簇中的被试编号，以及对应的实时连续情绪标注序列，进行单元层级和片段层级情绪融合。八个视频的标注数据融合

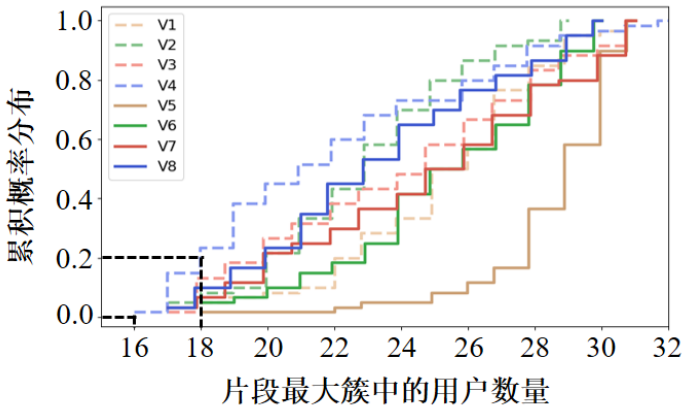


图 6.8　层次聚类的被试数量 CDF 分布情况

结果呈现在图6.9中。为了测试融合后的实时连续情绪标注数据的一致性和有效性，对每个视频的标注结果进行时序性分析。假设 A_{ij} 表示视频 $i \in [1, I]$ 第 $j \in [1, J]$ 个片段中被试唤醒维度标注值的融合结果，V_{ij} 表示效价标注值的融合结果；其中 I 和 J 分别表示视频总数量和片段总数量。将唤醒值和效价值分为“低，$(1, 5)$”和“高，$[5, 9)$”两个类别[79]，如果 $[A_{i1}, A_{i2}, \cdots, A_{iJ}]$ 序列中 50% 的唤醒值属于“低”或者“高”类别，将视频 i 在唤醒维度的预测值（Predicted Label）记为“低”或“高”标签；效价维度的预测值采用同样方法计算。

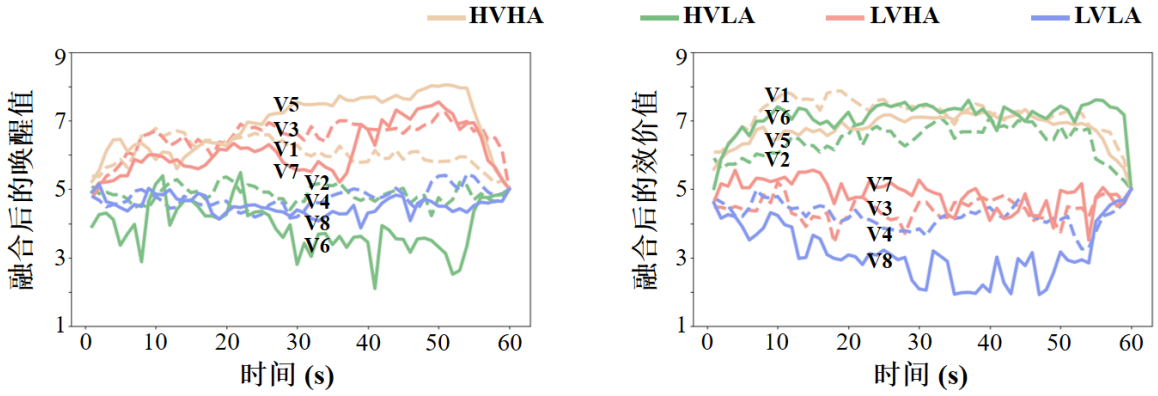


图 6.9　基于视觉交互行为的实时连续情绪融合结果

图6.10中的混淆矩阵展示了上述时序分析形成的唤醒预测值与效价预测值和诱发素材的原始情绪标签、以及体验后 SAM 标注结果（见3.4.1.3（4））之间的比较。这些矩阵表明融合后的唤醒和效价情绪标签能够精准地分类或预测原始标签以及 SAM 标签，分类准确率为 100%。

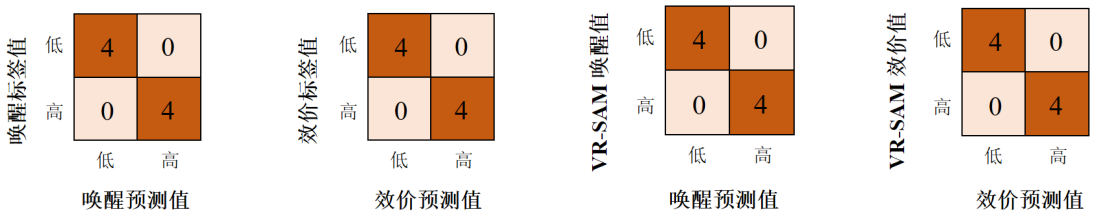
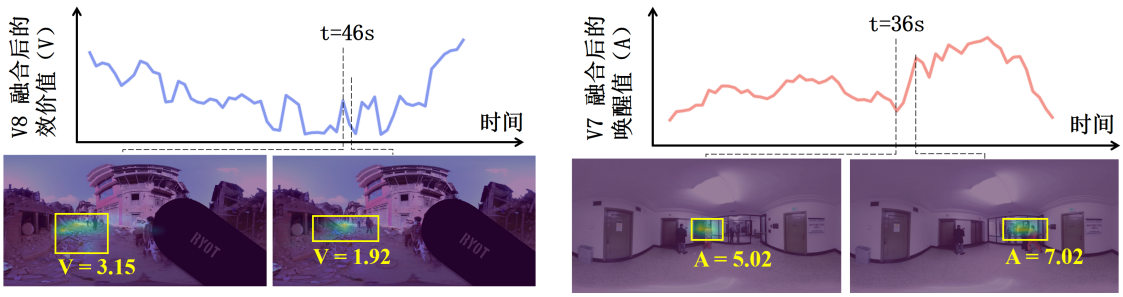


图 6.10　情绪预测值和原始情绪标签与 SAM 标注结果对比混淆矩阵

6.5.3　基于视觉交互行为的情绪识别讨论

本章提出的视口相关的情绪融合方法，能够实现在细粒度层级上理解并分析用户的情绪状态及对应的诱发情境。

针对 CEAP-360VR 数据集中的连续情绪融合结果，本小节首先讨论以下两个例子：（1）视频 V8 是一个讲述地震后场景的电影片段，视频第 10 秒左右的镜头为地震后的街道场景，情绪融合结果中效价值的变化平缓；但是在第 46 秒左右，一个大箱子突然从房顶掉落并扬起许多灰尘，效价值明显地从 3.15 降到 1.92，如图6.11(a)所示。（2）视频 V7 描述了一个罪犯在警察局越狱的片段，视频第 36 秒，一名狱警带领一名罪犯刚走出电梯口，罪犯突然挣脱并掉头逃跑，从图6.11(b)中可以看到融合结果中的唤醒值从 5.02 迅速升至 7.02。视频 V1 和 V5 的情绪融合结果在唤醒维度上也有相同的起伏。上述例子表明融合后的实时连续标注数据能够提供情绪状态的峰值、波谷、变化趋势等时序细节信息，有助于在细粒度层级分析情绪与对应诱发场景之间的相关联系。



(a) V8 融合后的效价值时序性分析　　(b) V7 融合后的唤醒值时序性分析

图 6.11　视口相关的情绪融合结果时序性分析案例

在虚拟环境中，由于用户情绪标注是基于视口连续变化的，这会造成在给定时间点用户之间情绪标注与对应场景的不确定性。现有技术通过特定视觉要素增强用户观看行为的一致性，例如，Liu 等人[250] 引入 Gated Clips 方法，当用户观看方向不在关键区域时，循环当前视口的内容，当用户视口位于关键区域时，继续视频内容播放，确保用户不会错过重要的叙事内容（Special Region of Interest, ROI）；Gruenefeld 等人[131] 设计了 HaloVR 与 WedgeVR 两种视觉线索引导用户的关注点。本研究旨在不更改诱发素材内容，在为用户提供尽可能自由地观看体验中获取更精确的情绪标签。实验结果验证了基于用户头部行为内在规律，构建视口相关的实时连续情绪标签方法的有效性和准确性，提供了在细粒度层级上分析情绪时序性特征的视角与工具。

值得注意的是，本实验将所有被试在同一诱发素材中的观看起始点校正为同一个位置（见3.4.2（4）），降低了被试在初始观看时视口位置的分散度。但是，在其他虚拟场景特别是自然环境中，用户体验的起始点校准会比较复杂，例如 6 自由度（6 Degree of Freedom，6DoF）视频、VR 游戏等，用户在体验过程中的头部运动行为模式也会更加分散、规律性更弱[212]。在这一层面上，本章提出的视口相关情绪融合方法会受到诱发素材类型的影响。随之而来的问题是：如果视口相关的聚类结果产生了两个或更多的“最大簇”（每个簇中的用户数量非常接近），是否需要对多个簇中的情绪标注数据分别进行融合分析；在更极端的情况下，用户的观看行为之间没有规律性，聚类后形成的簇数量与用户数量基本一致。在这些情况下，研究需要探索不同的体验初始点与视口聚类结果之间的关联、进行个性化的体验行为与情绪相关性分析。此外，这些情况的存在会影响开发用户独立的情绪识别模型的可行性[118,251]。然而，Jun 等人[33] 在最近一项包含 511 位被试的大规模研究中指出，用户喜爱度高的全景视频均具有能够吸引注意力的焦点，参与者在整体体验过程中对该类视频的探索较少。这证明了本章提出的视口相关的情绪融合方法用于更大规模的虚拟体验中的可行性和有效性；另一方面，这也意味着虚拟内容创作者可能需要在素材中定义视觉显著线索[252,253] 用于引导用户的视觉注意，从而改进视口聚类结果、提升视口相关的情绪融合的有效性和准确性。

6.6 本章小结

本章构建了虚拟环境中基于用户视觉交互行为的情绪识别方法，将用户的头部运动轨迹与实时连续情绪标注数据相结合，提出了连续情绪时间对齐及视口相关的情绪

融合方法，用于探索用户独立的情绪 Ground-Truth 标签。研究主要包括以下三个层面：（1）基于诱发素材参考特征的实时连续情绪标注序列时间对齐；（2）具有相似观看行为的用户片段层级视口聚类；（3）视口相关的情绪标注序列在单元层级和片段层级融合。研究采用上述方法对 CEAP-360VR 数据集中的实时连续情绪标注数据进行融合，并对融合结果展开分析。首先通过特征选择和相关性分析将诱发素材的颜色特征作为参考特征，计算 32 位被试观看八个全景视频时的情绪标注延迟时长；然后，根据用户在虚拟体验中的视觉行为进行片段层级视口聚类，结果表明用户在视频观看中具有较好的一致性；最后，对每个片段聚类结果最大簇中用户的情绪标注数据进行融合，得到的标签能够精准分类、预测诱发素材的原始标签及离散标签，并提供了情绪状态的峰值、波谷、变化趋势等时序细节信息，实现了在细粒度层级上理解并分析虚拟环境中用户的情绪状态与对应诱发情境之间的关联性。

本章为虚拟现实和情感计算领域的研究者提供了沉浸式全景视频观看体验中实时连续情绪标注数据的对齐与融合方法，有助于更好地理解和分析视口相关的情绪时序性特征。另一方面，基于视觉交互的情绪识别可以构建更精确的用户独立情绪 Ground-Truth 标签，用于情绪识别模型的训练与测试。

结论

全文总结

本文结合虚拟现实、人机交互和情感计算研究领域，分析了虚拟环境中用户交互行为及情绪识别研究现状。首先从理论层面对虚拟环境中的情绪识别概念模型、情绪识别系统及多模态、细粒度情绪识别方法进行研究，聚焦视觉交互行为与细粒度情绪识别。研究首次构建了虚拟交互体验中精确有效的连续情绪 Ground-Truth 标签及多模态情绪数据集，创造性地提出了面向视觉交互行为的细粒度情绪识别关键方法。本文的主要工作及创新点包括：

（1）构建了实时连续情绪标注方法及评估体系。针对虚拟交互环境中现有情绪标注方法不实时不连续、耗时长且干扰用户体验等问题，本文聚焦三个设计原则，采用高分辨率的 HMD 设备为用户提供高质量内容和带有摇杆的无线数字游戏控制器作为情绪标注设备；通过多领域专家共同设计，从情绪类型与强度等角度对标注方法进行多轮迭代评估，提出 HaloLight 和 DotSize 两种标注信息可视化方案。研究提出了一个连续情绪标注方法可用性评估体系，从虚拟交互环境中用户体验质量和标注数据的有效性两个方面给出了评估指标及对应的评估方法。研究首次构建了虚拟交互环境中实时连续情绪诱发及测量实验范式，评估基于 HaloLight 与 DotSize 的情绪诱发实验场景和情绪数据采集系统的可用性。实验结果表明两种实时连续情绪标注方法均能够收用户精确有效的唤醒和效价维度情绪 Ground-Truth 标签；在用户体验的晕动症、临场感和任务负荷方面没有显著性差异，且没有影响用户虚拟体验质量。该研究突破了在虚拟交互环境中进行实时连续情绪标注的关键技术。

（2）创建了一个虚拟环境中公开的多模态情绪数据集。数据收集是一个漫长、复杂且昂贵的过程，因此高质量数据集对于提升许多领域的研究和技术至关重要。本文公开了首个虚拟交互环境中实时连续生理及行为情绪标注多模态数据集（CEAP-360VR）[1]，包含 32 位被试观看八个全景视频的诱发素材实时采集的连续情绪数据、头部与眼部运动信息、瞳孔直径数据、外周生理信号、体验后问卷数据以及用户数据获取、处理与验证脚本。研究采用统计学方法从多个角度分别验证了多模态用户数据的有效性；采用机器学习技术进行一系列分类基线实验，进一步验证了 CEAP-360VR

[1]数据集链接：https://github.com/cwi-dis/CEAP-360VR-Dataset

数据集的有效性和可信度。实验表明 RF 分类器在 2 秒时长片段中分类性能较好，消融实验结果表明仅使用行为数据或生理信号均能够产生合理的识别准确度，但同时使用这两个类型的模态数据能够提升识别准确度。CEAP-360VR 数据集中的视觉行为数据可进一步探索视觉注意机制，多模态生理信号可用于探索隐式感知体验；为虚拟交互环境中的情绪识别研究提供了良好的数据源，极大地丰富了细粒度层级上用户情绪理解和预测研究。

（3）分析了视觉行为特征及其与连续情绪报告之间的细粒度相关性。针对虚拟环境中用户视觉交互行为的复杂性，本文主要关注用户佩戴 HMD 时的头部运动与眼部运动两个关键视觉行为要素，以及从中提取的注视、眼跳等行为特征值，探讨了虚拟环境中四种视觉交互行为特征：用户之间的头部运动与眼部运动一致性、用户头部运动与眼部运动之间的相关性、用户视觉行为的赤道及前方偏向、诱发素材内容对用户视觉行为的影响。为了进一步理解虚拟环境中视觉交互行为与情绪间的关系，研究首次提出了一种片段层级的用户头部运动、眼部运动与连续情绪标签之间相关性识别方法。实验发现虚拟体验中用户之间的视觉行为具有高一致性，视觉显著区域会受到观看内容影响，但仍具有明显的赤道和前方偏向；用户的头部运动与眼部运动及特征和实时连续的唤醒与效价情绪评分之间具有显著相关性。研究为根据用户视觉行为信息动态调整诱发内容奠定了基础，有助于改进虚拟内容处理、编码、传输和渲染技术。

（4）构建了基于视觉交互行为的连续情绪识别方法。本文首先在虚拟现实中情感计算研究领域引入了一个新问题——如何面向视觉交互行为特征构建用户独立的情绪 Ground-Truth 标签。为了解决这一问题，本文创造性地提出了基于视觉交互行为的情绪识别方法。针对用户个体之间的认知差异与反应延迟，建立了基于诱发素材参考特征的实时连续情绪标注序列时间对齐方法；针对虚拟体验中用户视觉交互行为的自由性与多样性，研究根据视觉行为模式进行片段层级视口聚类，提出了一种视口相关的情绪融合方法。实验结果显示，融合后的连续情绪标签能够精准预测诱发素材中原始标签和离散情绪标注标签。因此，本文提出的方法可用于构建虚拟交互环境中精确有效的 Ground Truth 情绪标签，同时还提供了情绪状态的峰值、波谷、变化趋势等时序细节信息，实现了在细粒度层级上理解并分析用户的情绪状态与对应诱发情境之间的关联性，推动了虚拟交互环境中细粒度情绪识别研究。

未来工作展望

本文的研究工作主要集中在虚拟交互环境中实时连续情绪 Ground-Truth 标签获取及面向视觉交互行为的细粒度情绪识别，并在实验中取得了积极的研究结果和认识。本文认为虚拟交互环境中的情绪识别及实际应用场景还存在诸多亟待解决的问题，需要进一步地思考和研究：

（1）复杂交互情境下的实时连续情绪标注。本文首次提出的实时连续情绪标注方法借助了外部摇杆设备，能够在不增加用户认知负荷的前提下有效地获取连续情绪 Ground-Truth 标签。但考虑到一些更为复杂或极端的虚拟交互情境中，例如用户在自由漫游状态下通过双手操纵手柄设备进行虚拟体验，这就需要开发新的情绪标注方式，以更自然的方式报告并收集情绪信号。因此，需要在保障用户体验质量、且降低标注方法学习成本的前提下，结合趋避范式等自然人机交互研究，探索如何通过人的某种模糊表达和传达方式，机器端能够精确地采集并理解情绪报告。此外，针对个体之间更复杂多样的行为特征，如何融合情绪标注序列、构建用户独立的连续情绪 Ground-Truth 标签，也是未来工作的研究方向。

（2）细粒度情绪识别算法构建。缺乏多模态情绪数据集与连续情绪 Ground-Truth 标签是限制细粒度情绪识别的根本原因。为此本文创建了虚拟交互环境中第一个公开且带有连续情绪 Ground-Truth 标签的 VR 数据集。结合机器学习相关技术，未来该数据集可用于评估验证弱监督学习、回归等细粒度情绪识别算法。由于虚拟环境中连续情绪标注数据获取难度大，未来研究可以采用已有数据训练生成式对抗网络等模型，用于扩充情绪样本数量；开发迁移学习、少样本学习等算法，更好地从已有数据中学习有效知识并降低偏差，推进虚拟现实领域的自动情绪识别研究。

（3）结合脑电信号的细粒度情绪识别。人的情绪起源于若干具有调节与感知功能的大脑皮层区域，生理心理学的很多研究也指出脑电信号在不同频段上与人的情绪状态密切相关，是一个良好的情绪预测因子。本文的研究中没有收集和使用脑电信号，是因为考虑到 EEG 采集设备由线缆和附着在头部的一组传感器组成，在沉浸式虚拟环境中，佩戴 HMD 的同时穿戴脑电设备会给用户带来较大的不适感，干扰用户体验；同时用户在虚拟体验中的头部运动会给脑电信号的稳定性带来挑战。为了解决这些问题，虚拟现实领域的情绪研究开始关注轻量级脑电设备，如仅带有三个电极的 NeuroSky，或是在 HMD 设备中嵌入脑电传感器，旨在获取虚拟环境中更加稳定可信的脑电信号。基于此，研究下一步将尝试采集脑电数据，探索细粒度层级的情绪标注

信号、行为特征及脑电数据之间的关系，结合脑电信号开展虚拟环境中的细粒度情绪识别。

参考文献

[1] 汪成为. 灵境 (虚拟现实) 技术的理论, 实现及应用[M]. 中国: 清华大学出版社, 1996.

[2] Barathi S C, Proulx M, O'Neill E, et al. Affect recognition using psychophysiological correlates in high intensity vr exergaming[C]. Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. New York, NY, USA: Association for Computing Machinery, 2020: 1–15.

[3] Nordahl R, Nilsson N. Oxford handbooks: The sound of being there: Presence and interactive audio in immersive virtual reality[M]. United Kingdom: Oxford University Press, 2014.

[4] Yoon Y, Moon D, Chin S. Fine tactile representation of materials for virtual reality[J]. Journal of Sensors, 2020, 2020.

[5] Kerruish E. Arranging sensations: smell and taste in augmented and virtual reality[J]. The Senses and Society, 2019, 14(1): 31–45.

[6] Sermet Y, Demir I. Flood action vr: A virtual reality framework for disaster awareness and emergency response training[C]. ACM SIGGRAPH 2019 Posters. New York, NY, USA: Association for Computing Machinery, 2019.

[7] Yang L, Huang J, Feng T, et al. Gesture interaction in virtual reality[J]. Virtual Reality & Intelligent Hardware, 2019, 1(1): 84–112.

[8] Chagué S, Charbonnier C. Real virtuality: A multi-user immersive platform connecting real and virtual worlds[C]. VRIC '16: Proceedings of the 2016 Virtual Reality International Conference. New York, NY, USA: Association for Computing Machinery, 2016.

[9] De Paolis L T, De Luca V. The impact of the input interface in a virtual environment: the vive controller and the myo armband[J]. Virtual Reality, 2020, 24(3): 483–502.

[10] 张凤军, 戴国忠, 彭晓兰. 虚拟现实的人机交互综述[J]. 中国科学（信息科学）, 2016, 46(12): 1711–1736.

[11] Picard R W. Affective computing[M]. United States of America: MIT press, 1997.

[12] Minsky M. Society of mind[M]. United States of America: Simon and Schuster, 1988.

[13] El Ali A, Perusquia-Hernandez M, Hassib M, et al. Meec: Second workshop on momentary emotion elicitation and capture[C]. Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems. New York, NY, USA: Association for Computing Machinery, 2021.

[14] Reisenzein R. Cognitive theory of emotion[J]. Encyclopedia of personality and individual differences, 2020: 723–733.

[15] Ruef A M, Levenson R W. Continuous measurement of emotion[J]. Handbook of emotion elicitation and assessment, 2007: 286–297.

[16] Gunes H, Schuller B. Categorical and dimensional affect analysis in continuous input: Current trends and future directions[J]. Image and Vision Computing, 2013, 31(2): 120–136.

[17] Shu L, Xie J, Yang M, et al. A review of emotion recognition using physiological signals[J]. Sensors, 2018, 18(7): 2074.

[18] Karg M, Samadani A, Gorbet R, et al. Body movements for affective expression: A survey of automatic recognition and generation[J]. IEEE Transactions on Affective Computing, 2013, 4(4): 341–359.

[19] Huang Z, Dang T, Cummins N, et al. An investigation of annotation delay compensation and output-associative fusion for multimodal continuous emotion prediction[C]. AVEC '15: Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge. New York, NY, USA: Association for Computing Machinery, 2015: 41–48.

[20] Metallinou A, Narayanan S. Annotation and processing of continuous emotional attributes: Challenges and opportunities[C]. 2013 10th IEEE international conference and workshops on automatic face and gesture recognition (FG). Shanghai, China: IEEE, 2013: 1–8.

[21] PS S, Mahalakshmi G. Emotion models: a review[J]. International Journal of Control Theory and Applications, 2017, 10: 651–657.

[22] Ekman P, Friesen W, Hager J. A technique for the measurement of facial action[J]. Palo alto, 1978.

[23] Plutchik R. The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice[J]. American scientist, 2001, 89(4): 344–350.

[24] Russell J A. A circumplex model of affect.[J]. Journal of personality and social psychology, 1980, 39(6): 1161.

[25] Soleymani M, Asghari-Esfeden S, Fu Y, et al. Analysis of eeg signals and facial expressions for continuous emotion detection[J]. IEEE Transactions on Affective Computing, 2015, 7(1): 17–28.

[26] Eerola T, Vuoskoski J K. A comparison of the discrete and dimensional models of emotion in music [J]. Psychology of Music, 2011, 39(1): 18–49.

[27] Felnhofer A, Kothgassner O D, Schmidt M, et al. Is virtual reality emotionally arousing? investigating five emotion inducing virtual park scenarios[J]. International Journal of Human-Computer Studies, 2015, 82: 48 – 56.

[28] Oliveira T, Noriega P, Rebelo F, et al. Evaluation of the relationship between virtual environments and emotions[C]. International Conference on Applied Human Factors and Ergonomics. Cham:

Springer International Publishing, 2018: 71–82.

[29] Peng X, Huang J, Li L, et al. Beyond horror and fear: Exploring player experience invoked by emotional challenge in vr games[C]. CHI EA '19: Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems. New York, NY, USA: Association for Computing Machinery, 2019: 1–6.

[30] Bhagat K K, Liou W K, Chang C Y. A cost-effective interactive 3d virtual reality system applied to military live firing training[J]. Virtual Reality, 2016, 20(2): 127–140.

[31] Kavanagh S, Luxton-Reilly A, Wuensche B, et al. A systematic review of virtual reality in education [J]. Themes in Science and Technology Education, 2017, 10(2): 85–119.

[32] Assilmia F, Pai Y S, Okawa K, et al. In360: A 360-degree-video platform to change students preconceived notions on their career[C]. Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems. New York, NY, USA: Association for Computing Machinery, 2017: 2359–2365.

[33] Jun H, Miller M R, Herrera F, et al. Stimulus sampling with 360-videos: Examining head movements, arousal, presence, simulator sickness, and preference on a large sample of participants and videos[J]. IEEE Transactions on Affective Computing, 2020: 1–1.

[34] on Virtual Reality & 3D User Interfaces I C. Session: Emotion and cognition[EB/OL]. 2022. https://ieeevr.org/2022/program/papers/#emotion-and-cognition.html.

[35] Marín-Morales J, Higuera-Trujillo J L, Greco A, et al. Affective computing in virtual reality: emotion recognition from brain and heartbeat dynamics using wearable sensors[J]. Scientific reports, 2018, 8(1): 1–15.

[36] Vesisenaho M, Juntunen M, Johanna P, et al. Virtual reality in education: Focus on the role of emotions and physiological reactivity[J]. Journal For Virtual Worlds Research, 2019, 12(1).

[37] Jang D P, Kim I Y, Nam S W, et al. Analysis of physiological response to two virtual environments: driving and flying simulation[J]. CyberPsychology & Behavior, 2002, 5(1): 11–18.

[38] Tang W, Wu S, Vigier T, et al. Influence of emotions on eye behavior in omnidirectional content[C]. 2020 Twelfth International Conference on Quality of Multimedia Experience (QoMEX). Athlone, Ireland, Ireland: IEEE, 2020: 1–6.

[39] Barrett L F. Feelings or words? understanding the content in self-report ratings of experienced emotion.[J]. Journal of personality and social psychology, 2004, 87(2): 266.

[40] El Ali A, Perusquía-Hernández M, Denman P, et al. Meec: First workshop on momentary emotion elicitation and capture[C]. Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems. New York, NY, USA: Association for Computing Machinery, 2020: 1–8.

[41] Cipresso P, Giglioli I A C, Raya M A, et al. The past, present, and future of virtual and augmented reality research: a network and cluster analysis of the literature[J]. Frontiers in psychology, 2018: 2086.

[42] Zhang T, El Ali A, Wang C, et al. Cornet: Fine-grained emotion recognition for video watching using wearable physiological sensors[J]. Sensors, 2021, 21(1): 52.

[43] Lloréns R, Noé E, Colomer C, et al. Effectiveness, usability, and cost-benefit of a virtual reality–based telerehabilitation program for balance recovery after stroke: A randomized controlled trial [J]. Archives of physical medicine and rehabilitation, 2015, 96(3): 418–425.

[44] Oberhauser M, Dreyer D. A virtual reality flight simulator for human factors engineering[J]. Cognition, Technology & Work, 2017, 19(2): 263–277.

[45] Howard M C, Gutworth M B. A meta-analysis of virtual reality training programs for social skill development[J]. Computers & Education, 2020, 144: 103707.

[46] Prayag G, Hosany S, Odeh K. The role of tourists' emotional experiences and satisfaction in understanding behavioral intentions[J]. Journal of Destination Marketing & Management, 2013, 2(2): 118–127.

[47] Alcañiz M, Bigné E, Guixeres J. Virtual reality in marketing: a framework, review, and research agenda[J]. Frontiers in psychology, 2019: 1530.

[48] Dantec M, Mantel M, Lafraire J, et al. On the contribution of the senses to food emotional experience [J]. Food Quality and Preference, 2021, 92: 104120.

[49] Burdea G C, Coiffet P. Virtual reality technology[M]. Canada: John Wiley & Sons, 2003.

[50] Regazzoni D, Rizzi C, Vitali A. Virtual reality applications: guidelines to design natural user interface[C]. International Design Engineering Technical Conferences and Computers and Information in Engineering Conference: volume 51739. [S.l.]: American Society of Mechanical Engineers, 2018: V01BT02A029.

[51] Fremerey S, Singla A, Meseberg K, et al. Avtrack360: An open dataset and software recording people's head rotations watching 360° videos on an hmd[C]. Proceedings of the 9th ACM Multimedia Systems Conference. New York, NY, USA: Association for Computing Machinery, 2018: 403–408.

[52] Corbillon X, De Simone F, Simon G. 360-degree video head movement dataset[C]. Proceedings of the 8th ACM on Multimedia Systems Conference. New York, NY, USA: Association for Computing Machinery, 2017: 199–204.

[53] Al Zayer M, MacNeilage P, Folmer E. Virtual locomotion: a survey[J]. IEEE transactions on visualization and computer graphics, 2018, 26(6): 2315–2334.

[54] Slater M, Steed A, Usoh M. The virtual treadmill: A naturalistic metaphor for navigation in immersive virtual environments[C]. Virtual environments’95. Vienna: Springer, 1995: 135–148.

[55] Terziman L, Marchal M, Emily M, et al. Shake-your-head: Revisiting walking-in-place for desktop virtual reality[C]. Proceedings of the 17th ACM Symposium on Virtual Reality Software and Technology. New York, NY, USA: Association for Computing Machinery, 2010: 27–34.

[56] Shiraishi T, Nakayama M. Eye movements and viewer’s impressions in response to hmd-evoked head movements[C]. Proceedings of the Workshop on Communication by Gaze Interaction. New York, NY, USA: Association for Computing Machinery, 2018.

[57] Tabbaa L, Searle R, Bafti S M, et al. Vreed: Virtual reality emotion recognition dataset using eye tracking & physiological measures[J]. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 2021, 5(4): 1–20.

[58] Laurutis V P, Robinson D A. The vestibulo-ocular reflex during human saccadic eye movements. [J]. Journal of Physiology, 1986, 373(1): 209–233.

[59] Piumsomboon T, Lee G, Lindeman R W, et al. Exploring natural eye-gaze-based interaction for immersive virtual reality[C]. 2017 IEEE Symposium on 3D User Interfaces (3DUI). Los Angeles, CA, USA: IEEE, 2017: 36–39.

[60] Andrist S, Gleicher M, Mutlu B. Looking coordinated: Bidirectional gaze mechanisms for collaborative interaction with virtual characters[C]. Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems. New York, NY, USA: Association for Computing Machinery, 2017: 2571–2582.

[61] Noroozi F, Corneanu C A, Kamińska D, et al. Survey on emotional body gesture recognition[J]. IEEE transactions on affective computing, 2018, 12(2): 505–523.

[62] Zhang F, Chu S, Pan R, et al. Double hand-gesture interaction for walk-through in vr environment [C]. 2017 IEEE/ACIS 16th International Conference on Computer and Information Science (ICIS). Wuhan, China: IEEE, 2017: 539–544.

[63] Li B J, Bailenson J N, Pines A, et al. A public database of immersive vr videos with corresponding ratings of arousal, valence, and correlations between head movements and self report measures[J]. Frontiers in psychology, 2017, 8: 1–10.

[64] Gusai E, Bassano C, Solari F, et al. Interaction in an immersive collaborative virtual reality environment: a comparison between leap motion and htc controllers[C]. International Conference on Image Analysis and Processing. Cham: Springer, 2017: 290–300.

[65] Lee J, Sinclair M, Gonzalez-Franco M, et al. Torc: A virtual reality controller for in-hand high-dexterity finger interaction[C]. Proceedings of the 2019 CHI Conference on Human Factors in

Computing Systems. New York, USA: Association for Computing Machinery, 2019: 1–13.

[66] Koons D B, Sparrell C J, Thorisson K R. Integrating simultaneous input from speech, gaze, and hand gestures[M]. Readings in intelligent user interfaces. New York, NY, USA: Association for Computing Machinery, 1998: 53–64.

[67] Tarnowski P, Kołodziej M, Majkowski A, et al. Eye-tracking analysis for emotion recognition[J]. Computational Intelligence and Neuroscience, 2020, 2020.

[68] Lim J Z, Mountstephens J, Teo J. Emotion recognition using eye-tracking: taxonomy, review and current challenges[J]. Sensors, 2020, 20(8): 2384.

[69] Milk C. Ted talk: How virtual reality can create the ultimate empathy machine [EB/OL]. 2022. https://www.ted.com/talks/chris_milk_how_virtual_reality_can_create_the_ultimate_empathy_machine.html.

[70] Somarathna R, Bednarz T, Mohammadi G. Virtual reality for emotion elicitation–a review[J]. arXiv preprint arXiv:2111.04461, 2021.

[71] Dozio N, Marcolin F, Scurati G W, et al. A design methodology for affective virtual reality[J]. International Journal of Human-Computer Studies, 2022: 102791.

[72] Riva G, Mantovani F, Capideville C S, et al. Affective interactions using virtual reality: the link between presence and emotions[J]. Cyberpsychology & behavior, 2007, 10(1): 45–56.

[73] Naz A, Kopper R, McMahan R P, et al. Emotional qualities of vr space[C]. 2017 IEEE Virtual Reality (VR). 2017: 3–11.

[74] Chirico A, Ferrise F, Cordella L, et al. Designing awe in virtual reality: An experimental study[J]. Frontiers in psychology, 2018, 8: 2351.

[75] Hedblom M, Gunnarsson B, Iravani B, et al. Reduction of physiological stress by urban green space in a multisensory virtual experiment[J]. Scientific reports, 2019, 9(1): 1–11.

[76] Koelstra S, Muhl C, Soleymani M, et al. Deap: A database for emotion analysis; using physiological signals[J]. IEEE transactions on affective computing, 2011, 3(1): 18–31.

[77] Soleymani M, Lichtenauer J, Pun T, et al. A multimodal database for affect recognition and implicit tagging[J]. IEEE transactions on affective computing, 2011, 3(1): 42–55.

[78] Sharma K, Castellini C, van den Broek E L, et al. A dataset of continuous affect annotations and physiological signals for emotion analysis[J]. Scientific data, 2019, 6(1): 1–13.

[79] Zhang T, El Ali A, Wang C, et al. Rcea: Real-time, continuous emotion annotation for collecting precise mobile video ground truth labels[C]. Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. New York, NY, USA: Association for Computing Machinery, 2020:

1–15.

[80] Nasoz F, Lisetti C L, Vasilakos A V. Affectively intelligent and adaptive car interfaces[J]. Information Sciences, 2010, 180(20): 3817–3836.

[81] Bălan O, Moise G, Moldoveanu A, et al. An investigation of various machine and deep learning techniques applied in automatic fear level detection and acrophobia virtual therapy[J]. Sensors, 2020, 20(2).

[82] He L, Li H, Xue T, et al. Am i in the theater? usability study of live performance based virtual reality[C]. Proceedings of the 24th ACM Symposium on Virtual Reality Software and Technology. New York, NY, USA: Association for Computing Machinery, 2018: 1–11.

[83] Kim Y, Moon J, Sung N J, et al. Correlation between selected gait variables and emotion using virtual reality[J]. Journal of Ambient Intelligence and Humanized Computing, 2019: 1–8.

[84] Reichenberger J, Pfaller M, Mühlberger A. Gaze behavior in social fear conditioning: An eye-tracking study in virtual reality[J]. Frontiers in psychology, 2020: 35.

[85] Ekman P. An argument for basic emotions[J]. Cognition & emotion, 1992, 6(3-4): 169–200.

[86] Darwin C. The expression of the emotions in man and animals[M]. United States of America: University of Chicago press, 2015.

[87] Mehrabian A. Comparison of the pad and panas as models for describing emotions and for differentiating anxiety from depregssion[J]. Journal of Psychopatholoxgy and Behavioral Assessment, 1997, 19(4): 331–357.

[88] Fontaine J R, Scherer K R, Roesch E B, et al. The world of emotions is not two-dimensional[J]. Psychological science, 2007, 18(12): 1050–1057.

[89] Bota P J, Wang C, Fred A L, et al. A review, current challenges, and future possibilities on emotion recognition using machine learning and physiological signals[J]. IEEE Access, 2019, 7: 140990–141020.

[90] Liu Z, Xu A, Guo Y, et al. Seemo: A computational approach to see emotions[C]. Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems. New York, NY, USA: Association for Computing Machinery, 2018: 1–12.

[91] Ringeval F, Sonderegger A, Sauer J, et al. Introducing the recola multimodal corpus of remote collaborative and affective interactions[C]. 2013 10th IEEE international conference and workshops on automatic face and gesture recognition (FG). Shanghai, China: IEEE, 2013: 1–8.

[92] McKeown G, Valstar M, Cowie R, et al. The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent[J]. IEEE transactions on affective computing, 2011, 3(1): 5–17.

[93] Bradley M M, Lang P J. Measuring emotion: the self-assessment manikin and the semantic differential[J]. Journal of behavior therapy and experimental psychiatry, 1994, 25(1): 49–59.

[94] Cowie R, Douglas-Cowie E, Savvidou S, et al. 'feeltrace': An instrument for recording perceived emotion in real time[C]. ISCA tutorial and research workshop (ITRW) on speech and emotion. Newcastle, Northern Ireland, UK: ISCA, 2000.

[95] Girard J M, Wright A G. Darma: Software for dual axis rating and media annotation[J]. Behavior research methods, 2018, 50(3): 902–909.

[96] Marín-Morales J, Llinares C, Guixeres J, et al. Emotion recognition in immersive virtual reality: From statistics to affective computing[J]. Sensors, 2020, 20(18): 5163.

[97] He C, Yao Y j, Ye X s. An emotion recognition system based on physiological signals obtained by wearable sensors[M]. Wearable sensors and robots. Cham: Springer, 2017: 15–25.

[98] Rigas G, Katsis C D, Ganiatsas G, et al. A user independent, biosignal based, emotion recognition method[C]. International Conference on User Modeling. Cham: Springer, 2007: 314–318.

[99] Wickramasuriya D S, Faghih R T. Online and offline anger detection via electromyography analysis [C]. 2017 IEEE Healthcare Innovations and Point of Care Technologies (HI-POCT). Bethesda, MD, USA: IEEE, 2017: 52–55.

[100] Chen L, Li M, Su W, et al. Adaptive feature selection-based adaboost-knn with direct optimization for dynamic emotion recognition in human–robot interaction[J]. IEEE Transactions on Emerging Topics in Computational Intelligence, 2019.

[101] Ma J, Tang H, Zheng W L, et al. Emotion recognition using multimodal residual lstm network [C]. Proceedings of the 27th ACM international conference on multimedia. New York, NY, USA: Association for Computing Machinery, 2019: 176–183.

[102] Ebrahimi Kahou S, Michalski V, Konda K, et al. Recurrent neural networks for emotion recognition in video[C]. Proceedings of the 2015 ACM on international conference on multimodal interaction. New York, NY, USA: Association for Computing Machinery, 2015: 467–474.

[103] D'mello S K, Kory J. A review and meta-analysis of multimodal affect detection systems[J]. ACM computing surveys (CSUR), 2015, 47(3): 1–36.

[104] Liu W, Zheng W L, Lu B L. Emotion recognition using multimodal deep learning[C]. International conference on neural information processing. Cham: Springer, 2016: 521–529.

[105] Wu D, Zhang J, Zhao Q. Multimodal fused emotion recognition about expression-eeg interaction and collaboration using deep learning[J]. IEEE Access, 2020, 8: 133180–133189.

[106] Hasanzadeh F, Annabestani M, Moghimi S. Continuous emotion recognition during music listening using eeg signals: A fuzzy parallel cascades model[J]. Applied Soft Computing, 2021, 101: 107028.

[107] Chang C Y, Zheng J Y, Wang C J. Based on support vector regression for emotion recognition using physiological signals[C]. The 2010 International Joint Conference on Neural Networks (IJCNN). Barcelona, Spain: IEEE, 2010: 1–7.

[108] Wei J, Chen T, Liu G, et al. Higher-order multivariable polynomial regression to estimate human affective states[J]. Scientific reports, 2016, 6(1): 1–13.

[109] Romeo L, Cavallo A, Pepa L, et al. Multiple instance learning for emotion recognition using physiological signals[J]. IEEE Transactions on Affective Computing, 2022, 13(1): 389–407.

[110] Zhang T, El Ali A, Hanjalic A, et al. Few-shot learning for fine-grained emotion recognition using physiological signals[J]. IEEE Transactions on Multimedia, 2022: 1–1.

[111] Awais M, Raza M, Singh N, et al. Lstm-based emotion detection using physiological signals: Iot framework for healthcare and distance learning in covid-19[J]. IEEE Internet of Things Journal, 2020, 8(23): 16863–16871.

[112] Srinivasan A, Abirami S, Divya N, et al. Intelligent child safety system using machine learning in iot devices[C]. 2020 5th International Conference on Computing, Communication and Security (ICCCS). 2020: 1–6.

[113] Oliveira T, Noriega P, Rebelo F, et al. Evaluation of the relationship between virtual environments and emotions[C]. Rebelo F, Soares M. Advances in Ergonomics in Design. Cham: Springer, 2018: 71–82.

[114] Nagel F, Kopiez R, Grewe O, et al. Emujoy: Software for continuous measurement of perceived emotions in music[J]. Behavior Research Methods, 2007, 39(2): 283–290.

[115] Doherty K, Doherty G. The construal of experience in hci: Understanding self-reports[J]. International Journal of Human-Computer Studies, 2018, 110: 63 – 74.

[116] Conner T S, Barrett L F. Trends in ambulatory self-report: The role of momentary experience in psychosomatic medicine[J]. Psychosomatic medicine, 2012, 74(4): 327.

[117] Toet A, Heijn F, Brouwer A M, et al. The emojigrid as an immersive self-report tool for the affective assessment of 360 vr videos[C]. International Conference on Virtual Reality and Augmented Reality. Cham: Springer, 2019: 330–335.

[118] Constantine L, Hajj H. A survey of ground-truth in emotion data annotation[C]. 2012 IEEE International Conference on Pervasive Computing and Communications Workshops. Lugano, Switzerland: IEEE, 2012: 697–702.

[119] Cowie R, McKeown G, Douglas-Cowie E. Tracing emotion: an overview[J]. International Journal of Synthetic Emotions (IJSE), 2012, 3(1): 1–17.

[120] Yannakakis G N, Martinez H P. Grounding truth via ordinal annotation[C]. 2015 international

conference on affective computing and intelligent interaction (ACII). Xi'an, China: IEEE, 2015: 574–580.

[121] Girard J M. Carma: Software for continuous affect rating and media annotation[J]. Journal of Open Research Software, 2014, 2(1).

[122] Lopes P, Yannakakis G N, Liapis A. Ranktrace: Relative and unbounded affect annotation[C]. 2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII). San Antonio, TX, USA: IEEE, 2017: 158–163.

[123] Chirico A, Gaggioli A. When virtual feels real: Comparing emotional responses and presence in virtual and natural environments[J]. Cyberpsychology, Behavior, and Social Networking, 2019, 22 (3): 220–226.

[124] Putze S, Alexandrovsky D, Putze F, et al. Breaking the experience: Effects of questionnaires in vr user studies[C]. Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. New York, NY, USA: Association for Computing Machinery, 2020: 1–15.

[125] Krüger C, Kojić T, Meier L, et al. Development and validation of pictographic scales for rapid assessment of affective states in virtual reality[J]. arXiv preprint arXiv:2004.00034, 2020.

[126] Voigt-Antons J N, Lehtonen E, Palacios A P, et al. Comparing emotional states induced by 360° videos via head-mounted display and computer screen[C]. 2020 Twelfth International Conference on Quality of Multimedia Experience (QoMEX). Athlone, Ireland: IEEE, 2020: 1–6.

[127] Bakker S, Hausen D, Selker T. Peripheral interaction: Challenges and opportunities for hci in the periphery of attention[M]. Cham: Springer, 2016.

[128] Matthews T, Dey A K, Mankoff J, et al. A toolkit for managing user attention in peripheral displays [C]. Proc. UIST '04. New York, NY, USA: Association for Computing Machinery, 2004: 247–256.

[129] Mairena A, Gutwin C, Cockburn A. Peripheral notifications in large displays: Effects of feature combination and task interference[C]. Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. New York, NY, USA: Association for Computing Machinery, 2019: 1–12.

[130] Gutwin C, Cockburn A, Coveney A. Peripheral popout: The influence of visual angle and stimulus intensity on popout effects[C]. Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems. New York, NY, USA: Association for Computing Machinery, 2017: 208–219.

[131] Gruenefeld U, Ali A E, Boll S, et al. Beyond halo and wedge: Visualizing out-of-view objects on head-mounted virtual and augmented reality devices[C]. Proceedings of the 20th International Conference on Human-Computer Interaction with Mobile Devices and Services. New York, NY, USA: Association for Computing Machinery, 2018.

[132] Jerald J. The vr book: Human-centered design for virtual reality[M]. New York, NY, USA: Association for Computing Machinery and Morgan & Claypool, 2015.

[133] Gigerenzer G. Why heuristics work[J]. Perspectives on psychological science, 2008, 3(1): 20–29.

[134] LaViola Jr J J. A discussion of cybersickness in virtual environments[J]. ACM Sigchi Bulletin, 2000, 32(1): 47–56.

[135] Sharma K, Castellini C, Stulp F, et al. Continuous, real-time emotion annotation: A novel joystick-based analysis framework[J]. IEEE Trans. Affective Computing, 2017.

[136] Kahneman D. Attention and effort: volume 1063[M]. United States of America: Prentice-Hall Englewood Cliffs, 1973.

[137] Wickens C D. Multiple resources and mental workload[J]. Human factors, 2008, 50(3): 449–455.

[138] Harrison B L, Ishii H, Vicente K J, et al. Transparent layered user interfaces: An evaluation of a display design to enhance focused and divided attention[C]. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. USA: ACM Press/Addison-Wesley Publishing Co., 1995: 317–324.

[139] Lindlbauer D, Lilija K, Walter R, et al. Influence of display transparency on background awareness and task performance[C]. Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems. New York, NY, USA: Association for Computing Machinery, 2016: 1705–1716.

[140] Ståhl A, Sundström P, Höök K. A foundation for emotional expressivity[C]. Proc. Designing for User Experience '05. United States of America: AIGA: American Institute of Graphic Arts, 2005: 33.

[141] Handayani D, Wahab A, Yaacob H. Recognition of emotions in video clips: the self-assessment manikin validation[J]. Telkomnika, 2015, 13(4): 1343.

[142] Norman D A, Draper S W. User centered system design; new perspectives on human-computer interaction[M]. USA: L. Erlbaum Associates Inc., 1986.

[143] Sanders E B N, Stappers P J. Co-creation and the new landscapes of design[J]. CoDesign, 2008, 4 (1): 5–18.

[144] Gibbs G R. Thematic coding and categorizing[J]. Analyzing qualitative data, 2007, 703: 38–56.

[145] Chattha U A, Janjua U I, Anwar F, et al. Motion sickness in virtual reality: An empirical evaluation [J]. IEEE Access, 2020, 8: 130486–130499.

[146] Bessa M, Melo M, Narciso D, et al. Does 3d 360 video enhance user's vr experience? an evaluation study[C]. Proceedings of the XVII International Conference on Human Computer Interaction. New York, NY, USA: Association for Computing Machinery, 2016.

[147] Reason J T, Brand J J. Motion sickness[M]. Oxford, England: Academic press, 1975.

[148] Kennedy R S, Lane N E, Berbaum K S, et al. Simulator sickness questionnaire: An enhanced method for quantifying simulator sickness[J]. The international journal of aviation psychology, 1993, 3(3): 203–220.

[149] Sheridan T B. Musings on telepresence and virtual presence[J]. Presence: Teleoperators & Virtual Environments, 1992, 1(1): 120–126.

[150] Schubert T, Friedmann F, Regenbrecht H. The experience of presence: Factor analytic insights[J]. Presence: Teleoperators & Virtual Environments, 2001, 10(3): 266–281.

[151] Hart S G. Nasa-task load index (nasa-tlx); 20 years later[C]. Proceedings of the human factors and ergonomics society annual meeting: volume 50. Los Angeles, CA: Sage Publications Sage CA, 2006: 904–908.

[152] Bradley M M, Miccoli L, Escrig M A, et al. The pupil as a measure of emotional arousal and autonomic activation[J]. Psychophysiology, 2008, 45(4): 602–607.

[153] Li B J, Bailenson J N, Pines A, et al. A public database of immersive vr videos with corresponding ratings of arousal, valence, and correlations between head movements and self report measures[J]. Frontiers in psychology, 2017, 8: 2116.

[154] Broeck M V d, Kawsar F, Schöning J. It's all around you: Exploring 360 video viewing experiences on mobile devices[C]. Proceedings of the 25th ACM international conference on Multimedia. New York, NY, USA: Association for Computing Machinery, 2017: 762–768.

[155] Lo W C, Fan C L, Lee J, et al. 360 video viewing dataset in head-mounted virtual reality[C]. Proceedings of the 8th ACM on Multimedia Systems Conference. New York, NY, USA: Association for Computing Machinery, 2017: 211–216.

[156] Hallgren K A. Computing inter-rater reliability for observational data: an overview and tutorial[J]. Tutorials in quantitative methods for psychology, 2012, 8(1): 23.

[157] Cicchetti D V. Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology.[J]. Psychological assessment, 1994, 6(4): 284.

[158] Recommendation I. 910, "subjective video quality assessment methods for multimedia applications," recommendation itu-t p. 910[J]. ITU Telecom. Standardization Sector of ITU, 1999.

[159] Zhao B, Wang Z, Yu Z, et al. Emotionsense: Emotion recognition based on wearable wristband[C]. 2018 IEEE SmartWorld, Ubiquitous Intelligence Computing, Advanced Trusted Computing, Scalable Computing Communications, Cloud Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI). Guangzhou, China: IEEE, 2018: 346–355.

[160] Gunst R F, Mason R L. Fractional factorial design[J]. Wiley Interdisciplinary Reviews: Computational Statistics, 2009, 1(2): 234–244.

[161] Lutz A, Brefczynski-Lewis J, Johnstone T, et al. Regulation of the neural circuitry of emotion by compassion meditation: effects of meditative expertise[J]. PloS one, 2008, 3(3).

[162] Fong C. Analytical methods for squaring the disc[J]. arXiv preprint arXiv:1509.06344, 2015.

[163] Fleureau J, Guillotel P, Orlac I. Affective benchmarking of movies based on the physiological responses of a real audience[C]. 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction. Geneva, Switzerland: IEEE, 2013: 73–78.

[164] Schwind V, Knierim P, Haas N, et al. Using presence questionnaires in virtual reality[C]. New York, NY, USA: Association for Computing Machinery, 2019: 1–12.

[165] Singla A, Fremerey S, Robitza W, et al. Measuring and comparing qoe and simulator sickness of omnidirectional videos in different head mounted displays[C]. 2017 Ninth international conference on quality of multimedia experience (QoMEX). Erfurt, Germany: IEEE, 2017: 1–6.

[166] Subramanyam S, Li J, Viola I, et al. Comparing the quality of highly realistic digital humans in 3dof and 6dof: A volumetric video case study[C]. 2020 IEEE Conference on Virtual Reality and 3D User Interfaces (VR). Atlanta, GA, USA: IEEE, 2020: 127–136.

[167] Pfleging B, Fekety D K, Schmidt A, et al. A model relating pupil diameter to mental workload and lighting conditions[C]. Proceedings of the 2016 CHI conference on human factors in computing systems. New York, NY, USA: Association for Computing Machinery, 2016: 5776–5788.

[168] Zhu Z, Fujimura K, Ji Q. Real-time eye detection and tracking under various light conditions[C]. Proceedings of the 2002 Symposium on Eye Tracking Research & Applications. New York, NY, USA: Association for Computing Machinery, 2002: 139‑144.

[169] Zhao S, Wang S, Soleymani M, et al. Affective computing for large-scale heterogeneous multimedia data: A survey[J]. ACM Trans. Multimedia Comput. Commun. Appl., 2019, 15(3s).

[170] MacQuarrie A, Steed A. Cinematic virtual reality: Evaluating the effect of display type on the viewing experience for panoramic video[C]. 2017 IEEE Virtual Reality (VR). Los Angeles, CA, USA: IEEE, 2017: 45–54.

[171] Xu M, Li C, Zhang S, et al. State-of-the-art in 360 video/image processing: Perception, assessment and compression[J]. IEEE Journal of Selected Topics in Signal Processing, 2020, 14(1): 5–26.

[172] Bian Y, Yang C, Gao F, et al. A framework for physiological indicators of flow in vr games: construction and preliminary evaluation[J]. Personal and Ubiquitous Computing, 2016, 20(5): 821–832.

[173] Egan D, Brennan S, Barrett J, et al. An evaluation of heart rate and electrodermal activity as an

objective qoe evaluation method for immersive virtual reality environments[C]. 2016 Eighth International Conference on Quality of Multimedia Experience (QoMEX). Lisbon, Portugal: IEEE, 2016: 1–6.

[174] Eudave L, Valencia M. Physiological response while driving in an immersive virtual environment [C]. 2017 IEEE 14th International Conference on Wearable and Implantable Body Sensor Networks (BSN). Eindhoven, Netherlands: IEEE, 2017: 145–148.

[175] Liao D, Shu L, Liang G, et al. Design and evaluation of affective virtual reality system based on multimodal physiological signals and self-assessment manikin[J]. IEEE Journal of Electromagnetics, RF and Microwaves in Medicine and Biology, 2019.

[176] Slater M, McCarthy J, Maringelli F. The influence of body movement on subjective presence in virtual environments[J]. Human factors, 1998, 40(3): 469–477.

[177] Won A S, Perone B, Friend M, et al. Identifying anxiety through tracked head movements in a virtual classroom[J]. Cyberpsychology, Behavior, and Social Networking, 2016, 19(6): 380–387.

[178] Livingstone S R, Palmer C. Head movements encode emotions during speech and song.[J]. Emotion, 2016, 16(3): 365.

[179] David E J, Gutiérrez J, Coutrot A, et al. A dataset of head and eye movements for 360 videos[C]. Proceedings of the 9th ACM Multimedia Systems Conference. New York, NY, USA: Association for Computing Machinery, 2018: 432–437.

[180] Li C, Xu M, Du X, et al. Bridge the gap between vqa and human behavior on omnidirectional video: A large-scale dataset and a deep learning model[C]. Proceedings of the 26th ACM international conference on Multimedia. New York, NY, USA: Association for Computing Machinery, 2018: 932–940.

[181] Miranda-Correa J A, Abadi M K, Sebe N, et al. Amigos: A dataset for affect, personality and mood research on individuals and groups[J]. IEEE Transactions on Affective Computing, 2018, 12(2): 479–493.

[182] Subramanian R, Wache J, Abadi M K, et al. Ascertain: Emotion and personality recognition using commercial sensors[J]. IEEE Transactions on Affective Computing, 2016, 9(2): 147–160.

[183] Bray T, et al. The javascript object notation (json) data interchange format[M]. [S.l.]: RFC 7159, DOI 10.17487/RFC7159, March 2014,< http://www. rfc-editor. org>, 2014.

[184] Partala T, Surakka V. Pupil size variation as an indication of affective processing[J]. International journal of human-computer studies, 2003, 59(1-2): 185–198.

[185] Kun A L, Palinko O, Razumenić I. Exploring the effects of size and luminance of visual targets on the pupillary light reflex[C]. Proceedings of the 4th International Conference on Automotive User

Interfaces and Interactive Vehicular Applications. New York, NY, USA: Association for Computing Machinery, 2012: 183–186.

[186] Wagner J, Kim J, André E. From physiological signals to emotions: Implementing and comparing selected methods for feature extraction and classification[C]. 2005 IEEE international conference on multimedia and expo. Amsterdam, Netherlands: IEEE, 2005: 940–943.

[187] Nabian M, Yin Y, Wormwood J, et al. An open-source feature extraction tool for the analysis of peripheral physiological data[J]. IEEE journal of translational engineering in health and medicine, 2018, 6: 1–11.

[188] Boucsein W. Electrodermal activity[M]. Cham: Springer Science & Business Media, 2012.

[189] Schmidhuber J. Deep learning in neural networks: An overview[J]. Neural networks, 2015, 61: 85–117.

[190] Kardan O, Berman M G, Yourganov G, et al. Classifying mental states from eye movements during scene viewing.[J]. Journal of Experimental Psychology: Human Perception and Performance, 2015, 41(6): 1502.

[191] Zou F, Shen L, Jie Z, et al. A sufficient condition for convergences of adam and rmsprop[C]. Proceedings of the IEEE conference on computer vision and pattern recognition. Long Beach, CA, USA: IEEE, 2019: 11127–11135.

[192] Fatourechi M, Ward R K, Mason S G, et al. Comparison of evaluation metrics in classification applications with imbalanced datasets[C]. 2008 seventh international conference on machine learning and applications. San Diego, CA, USA: IEEE, 2008: 777–782.

[193] Ekman P. Emotions revealed. second edition: Recognizing faces and feelings to improve communication and emotional life[M]. New York, NY, USA: OWL Books, 2007.

[194] Mirza M, Osindero S. Conditional generative adversarial nets[J]. arXiv preprint arXiv:1411.1784, 2014.

[195] Gutiérrez J, David E J, Coutrot A, et al. Introducing un salient360! benchmark: A platform for evaluating visual attention models for 360 contents[C]. 2018 Tenth International Conference on Quality of Multimedia Experience (QoMEX). Cagliari: IEEE, 2018: 1–3.

[196] Schouwstra S J, Hoogstraten J. Head position and spinal position as determinants of perceived emotional state[J]. Perceptual and motor skills, 1995, 81(2): 673–674.

[197] Ekman P, Friesen W V. Head and body cues in the judgment of emotion: A reformulation.[C]. Perceptual and Motor Skills: volume 246. Los Angeles, CA: Sage Publications, 1967: 711–724.

[198] De Lemos J, Sadeghnia G R, Ólafsdóttir Í, et al. Measuring emotions using eye tracking[C]. Proceedings of measuring behavior: volume 226. (Maastricht, The Netherlands: Proceedings of mea-

suring behavior, 2008: 225–226.

[199] Yarbus A L. Eye movements and vision[M]. New York, USA: Plenum Press, 1967.

[200] Ozcinar C, Smolic A. Visual attention in omnidirectional video for virtual reality applications[C]. 2018 Tenth International Conference on Quality of Multimedia Experience (QoMEX). Cagliari, Italy: IEEE, 2018: 1–6.

[201] Xu M, Song Y, Wang J, et al. Predicting head movement in panoramic video: A deep reinforcement learning approach[J]. IEEE transactions on pattern analysis and machine intelligence, 2018, 41(11): 2693–2708.

[202] Rai Y, Gutiérrez J, Le Callet P. A dataset of head and eye movements for 360 degree images[C]. Proceedings of the 8th ACM on Multimedia Systems Conference. New York, NY, USA: Association for Computing Machinery, 2017: 205–210.

[203] Xu Y, Dong Y, Wu J, et al. Gaze prediction in dynamic 360 immersive videos[C]. proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT, USA: IEEE, 2018: 5333–5342.

[204] Nguyen A, Yan Z, Nahrstedt K. Your attention is unique: Detecting 360-degree video saliency in head-mounted display for head movement prediction[C]. Proceedings of the 26th ACM international conference on Multimedia. New York, NY, USA: Association for Computing Machinery, 2018: 1190–1198.

[205] Adams A, Mahmoud M, Baltrušaitis T, et al. Decoupling facial expressions and head motions in complex emotions[C]. 2015 International Conference on Affective Computing and Intelligent Interaction (ACII). Xi'an, China: IEEE, 2015: 274–280.

[206] Aviezer H, Trope Y, Todorov A. Body cues, not facial expressions, discriminate between intense positive and negative emotions[J]. Science, 2012, 338: 1225 – 1229.

[207] Lhommet M, Marsella S C. Expressing emotion through posture[J]. The Oxford handbook of affective computing, 2014, 273: 1085–1101.

[208] Gross M M, Crane E A, Fredrickson B L. Methodology for assessing bodily expression of emotion [J]. Journal of Nonverbal Behavior, 2010, 34(4): 223–248.

[209] Wiebe A, Kersting A, Suslow T. Deployment of attention to emotional pictures varies as a function of externally-oriented thinking: An eye tracking investigation[J]. Journal of behavior therapy and experimental psychiatry, 2017, 55: 1–5.

[210] Fang Y, Nakashima R, Matsumiya K, et al. Eye-head coordination for visual cognitive processing [J]. PloS one, 2015, 10(3).

[211] Salvucci D D, Goldberg J H. Identifying fixations and saccades in eye-tracking protocols[C]. Pro-

ceedings of the 2000 Symposium on Eye Tracking Research & Applications. New York, NY, USA: Association for Computing Machinery, 2000: 71–78.

[212] Sitzmann V, Serrano A, Pavel A, et al. Saliency in vr: How do people explore virtual environments? [J]. IEEE transactions on visualization and computer graphics, 2018, 24(4): 1633–1642.

[213] Xu M, Li C, Chen Z, et al. Assessing visual quality of omnidirectional videos[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2018, 29(12): 3516–3530.

[214] Rai Y, Le Callet P, Guillotel P. Which saliency weighting for omni directional image quality assessment?[C]. 2017 Ninth International Conference on Quality of Multimedia Experience (QoMEX). Erfurt, Germany: IEEE, 2017: 1–6.

[215] Lebreton P, Raake A. Gbvs360, bms360, prosal: Extending existing saliency prediction models from 2d to omnidirectional images[J]. Signal Processing Image Communication, 2018: S0923596518302406.

[216] Zhu Y, Zhai G, Min X. The prediction of head and eye movement for 360 degree images[J]. Signal Processing Image Communication, 2018: S0923596518304946.

[217] Lebreton P, Fremerey S, Raake A. V-bms360: A video extention to the bms360 image saliency model[C]. 2018 IEEE International Conference on Multimedia & Expo Workshops (ICMEW). San Diego, CA, USA: IEEE, 2018.

[218] Monroy R, Lutz S, Chalasani T, et al. Salnet360: Saliency maps for omni-directional images with cnn[J]. Signal Processing: Image Communication, 2018, 69: 26–34.

[219] Bindemann M. Scene and screen center bias early eye movements in scene viewing[J]. Vision research, 2010, 50(23): 2577–2587.

[220] Judd T, Ehinger K, Durand F, et al. Learning to predict where humans look[C]. 2009 IEEE 12th international conference on computer vision. Kyoto, Japan: IEEE, 2009: 2106–2113.

[221] Nuthmann A, Henderson J M. Object-based attentional selection in scene viewing[J]. Journal of vision, 2010, 10(8): 20–20.

[222] Suzuki T, Yamanaka T. Saliency map estimation for omni-directional image considering prior distributions[C]. 2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC). Miyazaki, Japan: IEEE, 2018.

[223] Akoglu H. User's guide to correlation coefficients[J]. Turkish Journal of Emergency Medicine, 2018, 18(3): 91–93.

[224] Smith G D, Ebrahim S. Data dredging, bias, or confounding: They can all get you into the bmj and the friday papers[EB/OL]. British Medical Journal Publishing Group, 2002. https://www.bmj.com/content/325/7378/1437.

[225] Benjamini Y, Yekutieli D. The control of the false discovery rate in multiple testing under dependency[J]. Ann. Stat., 2001, 29.

[226] Benjamini Y, Hochberg Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing[J]. Journal of the Royal Statistical Society. Series B: Methodological, 1995, 57 (1): 289–300.

[227] Calvo R, D’ Mello S, Gratch J, et al. Expressing emotion through posture and gesture[J]. The Oxford Handbook of Affective Computing, 2015: 273–285.

[228] Afzal S, Chen J, Ramakrishnan K. Characterization of 360-degree videos[C]. Proceedings of the Workshop on Virtual Reality and Augmented Reality Network. New York, NY, USA: Association for Computing Machinery, 2017: 1–6.

[229] Mariooryad S, Busso C. Correcting time-continuous emotional labels by modeling the reaction lag of evaluators[J]. IEEE Transactions on Affective Computing, 2014, 6(2): 97–108.

[230] Nicolle J, Rapp V, Bailly K, et al. Robust continuous prediction of human emotions using multiscale dynamic cues[C]. New York, NY, USA: Association for Computing Machinery, 2012: 501–508.

[231] Sakoe H, Chiba S. Dynamic programming algorithm optimization for spoken word recognition[J]. IEEE transactions on acoustics, speech, and signal processing, 1978, 26(1): 43–49.

[232] Nicolaou M A, Zafeiriou S, Pantic M. Correlated-spaces regression for learning continuous emotion dimensions[C]. Proceedings of the 21st ACM International Conference on Multimedia. New York, NY, USA: Association for Computing Machinery, 2013: 773–776.

[233] Marmitt G. Modeling in vr: Measuring the accuracy of predicted scanpaths[D]. South Carolina, USA: Clemson University, 2002.

[234] Wu C, Tan Z, Wang Z, et al. A dataset for exploring user behaviors in vr spherical video streaming [C]. Proceedings of the 8th ACM on Multimedia Systems Conference. New York, NY, USA: Association for Computing Machinery, 2017: 193–198.

[235] Rossi S, De Simone F, Frossard P, et al. Spherical clustering of users navigating 360 content[C]. ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Brighton, UK: IEEE, 2019: 4020–4024.

[236] Nasrabadi A T, Samiei A, Prakash R. Viewport prediction for 360° videos: a clustering approach[C]. Proceedings of the 30th ACM Workshop on Network and Operating Systems Support for Digital Audio and Video. New York, NY, USA: Association for Computing Machinery, 2020: 34–39.

[237] Likas A, Vlassis N, Verbeek J J. The global k-means clustering algorithm[J]. Pattern recognition, 2003, 36(2): 451–461.

[238] Johnson S C. Hierarchical clustering schemes[J]. Psychometrika, 1967, 32(3): 241–254.

[239] Su M C, Liu Y C. A hierarchical approach to art-like clustering algorithm[C]. Proceedings of the 2002 International Joint Conference on Neural Networks: volume 1. Honolulu, HI, USA: IEEE, 2002: 788–793.

[240] Luo R C, Lin M H, Scherp R S. Dynamic multi-sensor data fusion system for intelligent robots[J]. IEEE Journal on Robotics and Automation, 1988, 4(4): 386–396.

[241] Ma S, Si G, Yue W, et al. An online monitoring measure consistency computing algorithm by sliding window in multi-sensor system[C]. 2016 IEEE International Conference on Mechatronics and Automation. Harbin, China: IEEE, 2016: 2185–2190.

[242] Zhao S, Gao Y, Jiang X, et al. Exploring principles-of-art features for image emotion recognition [C]. Proceedings of the 22nd ACM international conference on Multimedia. New York, NY, USA: Association for Computing Machinery, 2014: 47–56.

[243] Solli M, Lenz R. Color emotions for image classification and retrieval[C]. Conference on Colour in Graphics, Imaging, and Vision: volume 2008. Terrassa, Spain: Society for Imaging Science and Technology, 2008: 367–371.

[244] Mohseni S A, Wu H R, Thom J A. Automatic recognition of human emotions induced by visual contents of digital images based on color histogram[C]. 2017 International Conference on Digital Image Computing: Techniques and Applications (DICTA). Sydney, NSW, Australia: IEEE, 2017: 1–8.

[245] Machajdik J, Hanbury A. Affective image classification using features inspired by psychology and art theory[C]. Proceedings of the 18th ACM international conference on Multimedia. New York, NY, USA: Association for Computing Machinery, 2010: 83–92.

[246] Stricker M A, Orengo M. Similarity of color images[C]. Storage and retrieval for image and video databases III: volume 2420. [S.l.]: International Society for Optics and Photonics, 1995: 381–392.

[247] Haralick R M, Shanmugam K, Dinstein I H. Textural features for image classification[J]. IEEE Transactions on systems, man, and cybernetics, 1973(6): 610–621.

[248] Canny J. A computational approach to edge detection[J]. IEEE Transactions on pattern analysis and machine intelligence, 1986(6): 679–698.

[249] Xie L, Xu Z, Ban Y, et al. 360probdash: Improving qoe of 360 video streaming using tile-based http adaptive streaming[C]. Proceedings of the 25th ACM international conference on Multimedia. New York, NY, USA: Association for Computing Machinery, 2017: 315–323.

[250] Liu S J, Agrawala M, DiVerdi S, et al. View-dependent video textures for 360° video[C]. Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology. New York, NY, USA: Association for Computing Machinery, 2019: 249?262.

[251] Kolodyazhniy V, Kreibig S D, Gross J J, et al. An affective computing approach to physiological emotion specificity: Toward subject-independent and stimulus-independent classification of film-induced emotions[J]. Psychophysiology, 2011, 48(7): 908–922.

[252] Lin Y C, Chang Y J, Hu H N, et al. Tell me where to look: Investigating ways for assisting focus in 360 video[C]. Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems. New York, NY, USA: Association for Computing Machinery, 2017: 2535–2545.

[253] Speicher M, Rosenberg C, Degraen D, et al. Exploring visual guidance in 360-degree videos[C]. Proceedings of the 2019 ACM International Conference on Interactive Experiences for TV and Online Video. New York, NY, USA: Association for Computing Machinery, 2019: 1–12.

攻读学位期间发表论文与研究成果清单

发表论文

[1] **T. Xue**, A. el Ali, T. Zhang, G. Ding, P. Cesar. CEAP-360VR: A Continuous Physiological and Behavioral Emotion Annotation Dataset for 360° Videos. IEEE Transactions on Multimedia.（**SCI 一区 TOP，IF=6.513，对应第四、五章**）

[2] **T. Xue**, A. el Ali, T. Zhang, G. Ding, P. Cesar. RCEA-360VR: Real-time, Continuous Emotion Annotation in 360° VR Videos for Collecting Precise Viewport-dependent Ground Truth Labels. Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems.（**CCF 推荐 A 类会议，对应第三、六章**）

[3] **T. Xue**, A. el Ali, G. Ding, P. Cesar. Investigating the Relationship between Momentary Emotion Self-reports and Head and Eye Movements in HMD-based 360° VR Video Watching. Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems.（**CCF 推荐 A 类会议，对应第五章**）

[4] **T. Xue**, S. Ghosh, G. Ding, A. el Ali, P. Cesar. Designing Real-time, Continuous Emotion Annotation Techniques for 360° VR Videos. Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems.（**CCF 推荐 A 类会议，对应第三章**）

[5] **T. Xue**, A. El Ali, G. Ding, P. Cesar. Annotation Tool for Precise Emotion Ground Truth Label Acquisition while Watching 360° VR Videos. 2020 IEEE International Conference on Artificial Intelligence and Virtual Reality.（**EI 会议**）

[6] **T. Xue**, J. Li, G. Chen, P. Cesar. A Social VR Clinic for Knee Arthritis Patients with Haptics. Proceedings of the 2020 ACM International Conference on Interactive Media Experiences.（**Best Demo Award**）

[7] L. He, H. Li, **T. Xue**, D. Sun, S. Zhu, and G. Ding. Am I in the theater? Usability study of live performance based virtual reality. Proceedings of the 24th ACM Symposium on Virtual Reality Software and Technology.（**EI 会议**）

[8] Y. Wu, L. Zhang, G. Ding, **T. Xue**, F. Zhang. Modeling of Performance Creative Eval-

uation Driven by Multimodal Affective Data. International Journal of Interactive Multimedia & Artificial Intelligence. （ **SCI 三区，IF=3.137**）

发明专利

[9] **第一发明人**. 一种直播方案仿真设计验证方法 [P]: 中国，ZL20200828974.4.

[10] **第一发明人**. 一种智能直播仿真预演系统 [P]: 中国，ZL202010834533.5.

[11] **第三发明人**. 剧场表演重建方法及装置 [P]: 中国，ZL201910300378.6.

获奖情况

[12] **第一完成人**. 第六届中国（国际）大学生“互联网 +”创新创业大赛. 国家级金奖.

[13] **第二完成人**. 第十二届“挑战杯”中国大学生创业计划竞赛. 国家级银奖.

[14] **第 14 完成人**. 2021 年度中国仿真学会科学技术一等奖.

参与项目

[15] 冬奥会开闭幕式大型表演智能化创编排演一体化服务平台关键技术. 国家重点研发计划.

[16] 基于扩展现实的高达成度协同学习空间构建技术研究. 国家自然科学基金（面上）.

致谢

六年的硕博求学生涯即将结束，离别在即，万般不舍。回望走过的岁月，这期间所有的梦想与迷茫，痛苦与拼搏，艰辛与孤独，黑暗中的负重前行，都已风吹云散。成长和蜕变的道路上固然充满着汗水与泪水，是这一路上何其幸运遇到的人们，给予我坚持的勇气。

先生之风，山高水长。首先，我要特别感谢我的导师丁刚毅教授。在科研中，丁老师专业的指导、敏锐的观点和开阔的视野让我一步步知道了身为一名博士生如何发现科学问题，解决科学问题；在实践中，丁老师提供的参与七十周年国庆、百年党庆、冬奥会开闭幕式等国家大型活动经历令我受益匪浅；除此之外，在出国深造、就业等问题中丁老师都为我提供了非常宝贵的建议和的充分的帮助。在北理工的这六年，最幸运的是有您作为明灯指引我前进。师恩似海，借此论文完成之际，向老师献上学生最衷心的感谢和最诚挚的敬意。

在硕博阶段的学习里，我还要感谢数字表演实验室老师们的帮助、鼓励，感谢春风化雨的李鹏老师、多才温暖的黄天羽老师、细致严谨的马建东老师、可爱可亲的吴羽琛老师、邻家姐姐般的梁栋老师，以及李立杰老师、张龙飞老师、金乾坤老师、邢莉老师、唐明湘老师等，因为有了你们，才成就了这个专业敬业、团结向上的团队。同时，感谢实验室的小伙伴们，每一场大型活动，难数清多少次的熬夜编码，多少次的反复修改，从纸上方案到仿真编排，从点位推演到多轮联排，感谢大家在繁忙科研项目生活中的相互帮助与支持，带给我最珍贵的校园记忆。庆幸一直在和这个最优秀的团队一起，一直有在做最酷的事情。

同时，感谢国家留学基金委的资助，感谢荷兰数学与计算机国家科学研究院的 Pablo Cesar 教授、Abdallah El Ali 博士为我提供科研上的指导和帮助，这些在我撰写博士论文的过程中起到了巨大的作用。感谢 DIS 小组的同学以及阿姆斯特丹的小伙伴们，在我进行为期一年的联合培养博士期间给予我学术上的启发和生活上的帮助。

求学之路上，家人的无条件支持是我勇敢闯荡的最大底气。父母一直都是我最坚实的后盾，亦是我前行的动力；感谢你们的言传身教，让我成为更好的人，养育之恩，无以为报，未来我也会不断努力，永远做父母的骄傲。感谢我的男朋友，带给我温暖的关心与陪伴，给予我莫大的包容和理解。

感谢诸位评委老师对我论文工作的指导和修正，感谢计算机学院和研究生院的老师们对毕业工作给予的支持与帮助。

以梦为马，不负韶华。最后，感谢自己读博路上的努力与坚持，永远能够保持一颗炙热的心去面对生活的重重困难，感恩所有相遇与陪伴。纵有疾风起，人生不言弃。

作者简介

薛彤

1994 年 8 月 14 日出生于山西省曲沃县。

2012 年 9 月考入中国传媒大学数字游戏设计（游戏设计技术方向）专业，2016 年 7 月本科毕业并获得工学学士学位。

2016 年 9 月保送北京理工大学数字表演专业，攻读硕士研究生。在 2017 年 9 月硕博连读攻读软件工程博士学位至今，师从丁刚毅教授。

2019 年 10 月——2020 年 10 月在荷兰数学与计算机国家科学院（Centrum Wiskunde & Informatica）进行联合培养博士项目，指导教师为 Pablo Cesar 和 Abdallah El Ali。