arXiv:2302.06806v1 [cs.HC] 14 Feb 2023

Anchorage：通过锚定事件对客户服务视频中的满意度进行可视分析

Kam Kw ai Wong、Xingbo Wang、Yong Wang、Jianben He、Rong Zhang 和 Huamin Q u

摘要——通过视频通信提供客户服务，为通过分析客户满意度进行质量管理带来了新机遇。然而，由于缺乏可靠的自我报告反馈，服务提供者难以充分评估客户服务，也面临着对多模态视频记录进行繁琐调查的问题。我们推出了*Anchorage*，这是一款可视分析系统，通过总结客户服务视频中的多模态行为特征来评估客户满意度，并揭示服务流程中的异常操作。我们利用语义明确的操作，将结构化事件理解引入视频，帮助服务提供者快速定位感兴趣的事件。*Anchorage* 支持从服务和操作层面全面评估客户满意度，并通过多方面的可视化视图，高效分析客户行为动态。我们通过案例研究和精心设计的用户研究，对*Anchorage* 进行了全面评估。结果表明，该系统在利用客户服务视频评估客户满意度方面具有有效性和可用性。我们发现，在评估客户满意度时引入事件上下文，可以提升评估效果，同时不影响标注精度。我们的方法可应用于同时采集无标注非结构化视频和序列记录的场景。

索引词—视频数据、视频可视化、客户满意度、可视分析。

1 引言

CUSTOMER 满意度（下文简称“满意度”）是一项重要的服务质量指标，与人们对品牌形象、忠诚度及转换行为的感知高度相关 [1]。多项国际标准（例如 ISO 10004 [2]）提倡监测满意度，这有助于各种规模的组织评估绩效并获得管理洞见。满意度通常通过直接方式衡量（例如自我报告调查 [3]），也可通过间接方式推断（例如投诉 [4]）。由于获取客户的直接反馈成本高昂，直接测量的样本量通常有限，从而降低了结果的可靠性。服务后的总结性评估也缺乏自发性，且容易受到认知偏差的影响，例如诱导性问题和峰终定律 [5]。因此，亟需能够高效、实时评估满意度的自动化方法。

由于代理能够在更少的时间和地点限制下服务更多客户，远程服务日益普及。以往关于满意度分析的研究主要聚焦于传统的数字服务渠道，例如短信 [6], [7] 和电话 [8], [9]。人们预计，这些渠道将逐渐转向基于视频的通信，因为后者能够提升用户体验，尤其是在客户服务领域 [10]。在这一转型过程中，视频数据的补充为评估满意度带来了新的机遇，但也带来了两项主要挑战。

首先，所收集的视频数据具有多模态特征，因此难以处理和理解。由于情绪与满意度高度相关，因此被广泛用于对满意度建模。因此，先前的分析 [8], [9], [11] 利用多模态情绪特征来推断满意度水平。然而，情绪对满意度维度 [3], [12]可能具有不同的含义和影响。例如，抱怨服务的客户表现出强烈的负面情绪和较低的满意度；但微笑并说“谢谢”的客户或许只是出于礼貌，因此难以推断其满意度。客户服务中的双向沟通场景进一步加剧了这一问题。仅从客户的角度分析满意度是不够的，因为代理的行为可能是引发情感反应的根源 [13]。我们需要探究如何将情绪特征与其他行为线索结合起来，以更有效地分析满意度。

其次，客户服务具有特征分布稀疏、事件上下文多样的特点。视频记录可能持续数分钟，但所需特征（例如面部表情）可能只持续几秒。从大型视频集合中识别此类细微且瞬时的细节十分繁琐。此外，从视频中推断事件上下文还需要额外的关注和领域专业知识。非结构化视频数据缺乏适当的分段方案，无法有效概述整个服务及其各个片段。评估满意度时很少考虑顺序与时间关系 [9],，因此研究满意度模式会产生额外成本。

为应对这些挑战，我们提出使用锚点，即，使用具有语义意义的事件来描述服务流程（操作锚点）和可观察的人类行为（行为锚点），以表示关键事件特征。操作锚点通过提供序列和时间上下文，将有序的事件理解引入视频中。行为锚点以紧凑且直观的方式表示多模态人类行为，从而实现自动化满意度评估。我们结合这两类锚点来定位感兴趣的事件，并结合上下文解读满意度水平的变化过程。

我们提出了 ***Anchorage***，这是一种可视分析系统，通过总结客户服务视频中的多模态行为特征，并揭示服务流程中的异常事件，来评估客户满意度。

• *KK Wong、X. Wang、J. He、R. Zhang* 和 *H. Qu* 均任职于香港科技大学。电子邮箱：{*kkwongar*、*xingbo.wang*、*jhebtr*、*zhangab*}*@connect.ust.hk* 以及 *huamin@cse.ust.hk*

• *Y. Wang* 任职于新加坡管理大学。电子邮箱：*yong- wang@smu.edu.sg*。他是通讯作者。

*M*稿件于*2005*年*4*月*19*日收到；修订于*2005*年*8*月*26*日，*2 015.*

Anchorage 基于多视角异常检测框架生成潜在的操作锚点和行为锚点，并提供初步的满意度估计。我们设计了一组协调联动的可视化，用于分析以锚点为背景的满意度，从而以更高的序列和时间分辨率呈现传统满意度评分。我们通过一项案例研究和一项精心设计的用户研究，验证了 *Anchorage* 的有效性。我们发现，在视频分析中引入事件上下文（即，锚点），能够提升满意度评估任务的表现，同时不损害标注精度。*Anchorage* 有助于总结视频内容、识别异常事件并理解多模态特征。我们的方法可应用于同时收集无标注非结构化视频和序列记录的场景。总而言之，我们的主要贡献如下：

- ◇通过与领域专家反复讨论，描述了利用客户服务视频和机器日志评估满意度水平的问题特征。我们将这些认识用于创建一个临时构建的数据集，以验证该方法在不同满意度场景下的有效性。◇提出了一个多层异常检测框架，用于生成锚点，从而提升视频分析中的事件理解效率；并将离散事件分析拓展至视频可视分析。◇ 设计了新颖且富有隐喻性的可视化方案，便于有效识别多模态异常事件，从而评估客户服务视频中的满意度水平。

2 相关工作

我们的工作与客户满意度分析、多模态视频的可视分析及其事件理解，以及离散事件序列可视化方面的既有研究相关。

2.1 数据驱动的客户满意度分析

研究人员已针对多种数据类型开展自动满意度评估，例如文本 [6], [7], [14], 、眼动轨迹 [15],和脑电图信号 [16]。这些方法依赖较高的主动参与率及特定设备，而这两者在许多服务场景中都并不常见。与其他数据源不同，视频更易获取，可提供连续输入以进行深入的满意度分析 [17], ，且不会给客户带来额外的认知负担 [18]。例如，研究人员分析了零售商店中的监控视频 [19] 和众包网络摄像头视频 [18], [20] ，以了解客户对产品和服务的反应。

许多基于视频的方法侧重于提取面部表情作为帧级特征，并评估整体视频级满意度 [21]–[23]。Yolcu 等人 [24] 汇总了不同客户的所有情绪，并将其作为估计客户满意度的替代指标。他们还考虑了头部姿态，以应对目标人物面部被遮挡时性能不佳的问题。然而，这些方法仅以视频中面部表情识别的准确度作为评估依据，并未评估满意度。它们只在情绪与满意度之间套用了一个简单的模型。

除了视觉通道，声学信息对于满意度估计也至关重要。Park 和 Gates [8] 为支持向量机提取了若干韵律和词汇特征，用于建模满意度。Seng 和 Ang [11] 采用专为满意度评估设计的线性模型，融合视觉和声学通道中的情感特征。Ando 等人 [9] 提出了一种分层框架，用于结合对话和单个话语中的情绪特征。这些方法为视频提供了数据驱动的满意度分数，并在真实数据集和临时构建的数据集上进行了测试。然而，视频级满意度评分无法识别关键转折点，也无法区分相互抵消案例。此外，这些方法很少或完全没有关注代理的行为，而这些行为可能是影响客户后续反应的先行事件 [1], [17]。我们的工作探索将行为特征与事件上下文进行多模态融合，以建立背景关联，从而实现交互式满意度分析。

2.2 多模态视频的视觉分析

学界已提出大量可视化方法，用于汇总和呈现视频数据 [25]–[30]。在许多应用问题中，通常会从视频中提取情绪特征。EmotionCues [25] 通过流图设计概述了课堂视频中的情感动态，并突出显示了模型不确定性。EmotionMap [26]和 E-ffective [29] 提出了基于地图和基于螺旋的设计，用于呈现多媒体视频中情感维度的时间概览。这些可视化有效呈现了重要时刻的视觉线索。然而，为满足各种特定领域的分析目的，视频通常需要纳入情绪以外的更多特征。

近期，人们开始使用更多模态，以补充视觉通道中缺失的信息。EmoCo [31]探究了面部表情、音频情感与转录文本情感之间的情感一致性，以减少对某一特定模态的过度依赖。Li 等人 [32] 将头部姿态与鼠标移动数据可视化，以关联在线监考中的多模态行为。VideoModerator [33] 从图像和转录文本中提取风险相关特征，以辅助直播审核。除了验证特征之外，分析多模态也有助于解读多模态模型 [34], [35], 、查询大型视频集合 [36],以及标注出声思维可用性测试视频 [37], [38]。尽管整合更多相关数据源能够提高分析结果的可信度 [39], [40], ，但不可避免地会增加理解这些结果时的认知负荷。我们采用一组协调良好的视图，便于直观解读多模态特征。具体而言，我们提出一种基于散点图的隐喻式可视化设计，用于概括多模态特征并展示满意度变化。我们利用机器日志为客户服务视频提供上下文信息，并帮助理解其流程。这为概括视频序列提供了新的视角。

2.3 视频可视分析中的事件理解

根据 H¨oferlin 等人 [41], 的观点，视频可视分析有三个主要目标：状态判定、事件检测和模型生成。状态判定用于识别帧级特征，例如目标跟踪。相比之下，另外两个目标关注视频中更大范围的内容。事件检测旨在定位特定事件发生的时刻，而模型生成则致力于从视频集合中挖掘常见模式，以便之后用于检测事件。我们的工作聚焦于模型生成，以理解未标注的满意度模式。

由于期望的模式定义模糊，分析人员必须探索庞大的低层事件空间，以生成高层概念 [37]。电影分析中的帧-镜头-场景层级 [42] 以及体育视频标注中的对象-事件-战术层级 [43] 可以结合起来，用以说明视频事件的复杂性和相互依赖性。为简化探索过程，EventAnchor [44]追踪球拍类运动视频中视觉上可见的对象，并将其状态的关键变化标记为

锚点。这些锚点被绘制在屏幕上，以指示对象的位置，从而实现对机器误差的交互式校准。然而，基于计算机视觉的锚点数量有限，无法轻易应用于声学通道及其他模态。此外，许多真实场景的视频记录中并不存在记分牌、场景变化等视觉效果，无法据此提供事件上下文。受锚点概念启发，我们拓展其在理解多模态视频事件方面的用途，并将其推广至视频与序列记录同步采集的场景。我们利用服务流程中语义明确的事件，优先安排对不同锚点的调查工作。我们提出了一种半自动框架，利用异常检测方法生成候选锚点，并通过直观的可视化设计支持候选项验证。

2.4 离散事件序列可视化

离散事件分析通常适用于日志数据，例如计算机系统记录 [45]，。这类记录中的每条事件记录都有时间戳，每种事件类型都有其语义含义。这些事件的多尺度时间结构常用于生成可视化摘要。例如，可以通过多变量正则表达式 [46] 和层次聚类 [47] 聚合事件序列，以满足不同的细节层次需求。分支区分 [48] 以及对事件序列中频繁模式的捆绑 [49]，有助于进一步理解事件演化模式中的分歧与汇聚。近期的一项综述 [50] 总结了事件序列的设计空间。

借助多尺度概览，分析人员可以直观地比较不同事件序列的时间特征，以定位异常行为 [51]和流程漂移 [52]。Guo 等人 [53]提出了一种基于变分自编码器的方法，用于检测事件的任意排序、缺失和重复。我们借鉴了这类研究的思路，以发现异常的服务流程。我们采用基于相似度的方法来查找时间异常，并使用马尔可夫技术检测序列异常。他们构建操作锚点，并结合多模态行为作为锚点的上下文来评估客户满意度。

3 问题特征描述

本节介绍满意度评估的背景
客户服务中。我们介绍用于评估的需求分析，以及临时构建的数据集的详细信息。

3.1 满意度与客户服务

客户满意度通常被广泛定义为客户期望与感知服务质量相符 [1], [2]。然而，正如“期望”和“感知”这两个高度主观的术语所暗示的那样，满意度是一个复杂的概念，不同的人对此有不同的理解。因此，通常会对服务进行录音，以防止错误信息并避免投诉中的纠纷，从而为分析提供丰富的录音资料。

客户服务是指组织在售前、售中和售后为客户提供的帮助与服务 [2]。例如，联络中心通过电话提供客户服务，以处理咨询、管理订单并排除故障 [10]。虽然已设立自助服务终端、移动应用程序和虚拟助手，让客户自行办理事务，但由于监管要求、业务流程复杂性，以及机器在引导人类意图方面能力不足，由员工协助提供的客户服务仍不可替代 [54]。

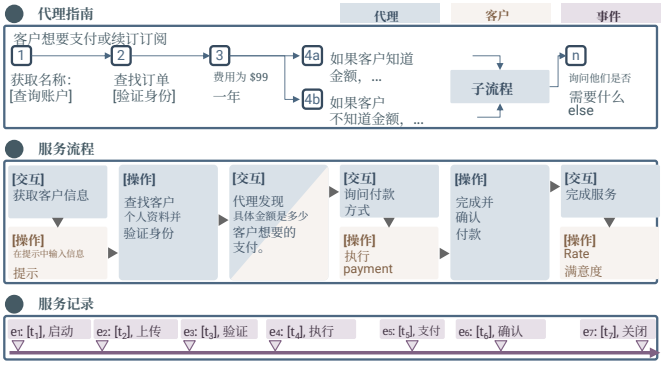


图 1. 改编自在线购物客户服务对话数据集的客户服务示例 [54]。(A) 代理指南概述了典型的实验流程。(B) 服务流程交织着代理与客户之间的交互和操作。(C) 服务记录会记录已执行的操作及其元数据，例如时间戳。

客户服务具有目标导向和沟通动态等特点。我们聚焦于一种典型情境：客服人员协助客户完成领域特定目标。我们在图 1 中概述了典型的工作流程，并阐释本文中使用的术语。首先，在服务开始时，代理协助客户明确其意图并理解其需求。其次，代理遵循内部指南，列出一系列操作，并引导客户完成顺序服务流程。第三，代理和客户通常轮流沟通并执行操作。例如，为了验证客户的身份，代理可以口头提问、提示客户输入信息，或查询数据库以获取个人资料。最后，已执行的操作构成服务记录，由于现实情况复杂，该记录可能与代理指南有所偏差。服务通常会录制为服务视频。

3.2 设计要求分析

我们采用设计研究方法 [55]来描述领域问题。在过去两年里，我们与一家国内信息技术公司合作，该公司致力于公共服务数字化，并为税务机关开发远程服务终端。这些终端通过视频通信连接身处不同地点的代理与客户，支持传输法律文件、处理数字支付等重要操作。设备会收集工作人员协助办理税务申报服务的视频记录 和机器日志 。

为了解其工作流程和满意度评估方法，我们对三位具有 5 年以上服务经验的一线税务官员进行了情境访谈和半结构化访谈（**E1-3**）。我们还采访了一位在客户关系领域拥有 10 年研究经验的教授（**E4**），以获取来自营销领域的另一种观点。在他们的影响下，我们阅读领域文献，了解重要概念，并将其与可视化研究联系起来。我们与行业合作方的三位业务分析师（**E5-7**）举行了一系列远程会议，他们均拥有 5 年以上公共服务经验。会议旨在完善设计要求，将先前研究发现应用于我们的场景，并验证迭代设计。**E1-7**均非本文的共同作者。

我们的目标是设计一套面向客户服务提供商的满意度评估系统，以识别满意度模式，从而改进其服务和工作流程。我们确定了六项设计要求，以支持*Anchorage*的开发。

R1 按客观指标对满意度进行排序。客户很少会在接受服务后提供满意度反馈。即使提供了，他们的自我报告评价也容易受到峰终定律等认知偏差的影响 [5] 尽管如此。**E1-2** 补充道：“有些客户急着离开，所以随便点了几个按钮。”大型视频集合也需要按优先级排序，以便优先查看感兴趣的视频。系统应根据用户行为提供统一且客观的评估。

R2 将满意度评价与操作情境化。应结合前置事件 [4] 来解读客户行为。例如，客户期待服务顺畅，因此会将反复且中断的操作视为麻烦且令人不满，从而产生消极的情绪反应。然而，客户对挑衅行为会有不同的情感反应。**E7** 指出：“有些人面无表情，但内心可能怒火中烧”，这表明仅使用情绪特征并不可靠，也不充分。系统应将流程因素纳入考量，作为解释和评估客户行为的共同基础。

R3 展示服务中的满意度变化。自动化方法通常会汇总帧级评价来建模满意度 [11]。然而，汇总后的服务评分难以区分相互抵消案例。**E5**提出了一个满意案例：客户起初表现出不满意行为，但到最后对服务更加满意。累积的服务级满意度评分无法充分体现这一案例。通过单项操作评估满意度，自然会凸显各项操作对总体评估的贡献 [9]。系统应将满意度变化动态可视化，以揭示行为与满意度之间的因果关系。

R4 概述服务记录。服务记录提供序列和时间信息，以表明服务流畅度。流畅的服务记录通常与代理指南中明确描述的典型工作流程相符。资深坐席（**E1-3**）阅读记录并理解其语义后，能够轻松识别偏离流程的操作。操作时长也会隐含地反映流程难点和服务状态。系统应提供充分的事件上下文，以增强流程意识并合理划分服务。

R5 突出显示异常操作。满意度通常会随着先前状态而稳步变化。显著的转折点可能表明，满意度模式受到内部因素（例如，超出预期 [1]）和外部因素（例如，坐席不当行为 [4]）的影响。**E4** 表示，关注满意度水平的“峰值”和“低谷”有助于获得更多管理洞见。这些情况值得进一步调查并给予更多关注。系统应区分异常的满意度变化，以识别关键转变时刻。

R6 支持对原始视频进行交互式导航。视频记录是评估满意度最有力的证据。然而，从头开始回看视频效率低下。机器学习模型提取的特征很有帮助，但可能受到模型不确定性和多模态交互的影响 [35]。**E6-7** 表示，当这些特征传达出“不合理且相互矛盾的含义”时，需要对其进行验证。此外，坐席与客户之间的互动动态难以定义和检测。系统应支持多种交互方式，以便快速定位感兴趣的事件。

3.3 临时构建的客户服务视频数据集

我们分析了来自**E5-7**的 20 段真实视频，内容涉及地方政府税务机关提供的税务服务协助。这些视频没有自报满意度水平。大多数客户面部表情中性，服务平均耗时。这些视频只能反映满意度中性的案例，因为**E6** 和**E7** 将其评估结果作为真实标注。由于新冠疫情封锁限制了我们前往地方税务办公室获取客户知情同意，因此很难获得更多录音。据我们所知，目前没有公开可用的数据集同时包含视频记录和服务记录，可用于评估满意度。为解决各满意度水平类别分布不平衡的问题，我们创建了一个用于概念验证的临时数据集。

参与者与实验装置。我们邀请了行业合作方的26名员工（4名女性和22名男性）参与情景演绎。这些参与者分别扮演代理或客户。四名业务分析师因熟悉服务工作流程而担任代理。其他参与者则担任客户，条件是他们曾前往政府机构办理工作人员辅助公共服务。我们使用合作方的终端设备进行录制，以模拟真实照明和遮挡环境。本研究已获内部机构审查委员会批准（#HREP- 2021年-0162），录制视频前均取得了书面同意。

设计与设置。 我们设计并演示了具有不同满意度水平的典型服务场景。通过观察收集到的视频片段中的工作流程，并采访一线服务人员（**E1-3**），我们归纳出四种满意度类型，包括：

ST 满意服务，完成时间短于预期。代理给出清晰的指示，并熟练地完成任务。客户表示感谢。**NM** 正常服务，完成时间符合预期。代理将每项操作的用时控制在平均水平左右。客户不会获得任何指示。**DA** 对代理不满意的服务，完成时间长于预期。代理表现出不专心行为（例如，使用手机、与他人闲聊），以延长服务时间。客户感到恼火。

对流程不满意：预期的完成时间更短，但实际完成时间更长，因此对**服务流**程不满意。服务流程因故障终端而中断，客户需要重复进行某些操作。客户感到恼火。

我们以较为概括的表述说明了各种满意度类型，作为指导。参与者自行设计对话和反应，即兴呈现服务过程。我们要求代理保持中性表情，以防止情绪传染 [13]。客户的反应则用“表示感谢”和“感到恼火”等抽象表述来描述。我们没有为这些场景编写脚本，以保持自发性、增强泛化能力，并避免知识诅咒。

服务场景围绕社会保障保险会员信息变更展开。历史记录显示，这项服务的平均耗时约为 8 分钟。由于预期会显著影响满意度 [1]，我们将这一信息告知参与者，作为预期时间，以控制所有参与者的时间预期。如图 3B1 所示，典型流程从启动到关闭，共包含九个步骤。我们的服务场景工作流程可迁移至一个面向客户服务的面向任务的数据集 [54]。

结果。 我们收集了61段服务视频，每种满意度类型至少有12段。我们确保同一客户不会重复扮演同一种类型。服务总时长为5.8小时，每段视频时长为3至12分钟，平均为5.7分钟。视频的真实标注为对应的满意度类型。

4 锚点生成

本节介绍锚点的构建。锚点是指具有语义意义的事件，用于描述操作（操作锚点）和行为（行为锚点）。我们采用多视角异常检测框架（**R5**），识别服务和操作中的异常事件，并将其作为操作锚点。我们进一步从服务视频中提取多模态特征，计算服务和操作的基本满意度估计，并将其作为行为锚点（**R1**）。这些锚点可视为交互式目录，用于定义视频事件结构（**R2**），以便快速定位满意度模式中所需片段，无需搜索整段视频（**R6**）。

4.1 处理多模态特征

我们分别处理了从视频中分解出的视觉和音频通道，并将其与解析后的机器日志对齐。

视觉特征。我们使用 YOLO5Face [56] 检测每一帧中的人脸边界框，并采用三角平滑来减少伪影。由于遮挡仍然是面部表情识别（FER） [57], 面临的挑战，我们使用 FSA-Net [58] 检测头部姿态，以验证其可靠性并降低误分类的影响。我们采用了一个面部表情识别模型 [59]，并将其输出的离散情绪汇总为三大类（积极、中性和消极），因为离散情绪与满意度之间的对应关系尚不明确 [1], [24]。**E5-7** 在评估客户服务满意度时，对悲伤和恐惧所起的作用也存在混淆。此外，为了提升泛化能力而牺牲粒度，是情感分析中的常见做法 [17]。

音频特征。 我们应用说话人分离算法 [60] 来去除噪声、定位语音片段，并根据说话人对话语进行聚类。由于视频中有两位说话人，我们通过启发式方法对代理进行注册，例如识别两个由同一代理处理的视频中的共同说话人。音频片段被输入音频情绪分类模型 [61]。离散情绪输出也会像面部表情一样进行汇总。

事件特征。 机器日志包含描述服务记录中操作的离散事件。然而，由于编码风格不一致，一个操作可能会由多条非结构化自由文本消息表示。

我们首先将日志转换为元组表示，其中包含时间戳、事件类型和日志参数列表，以供分析。通过匹配具有相同终端请求 ID 的起始和结束日志消息来识别一项服务。我们汇总了语义相关且共同出现的原始事件类型，并通过**E5**确认汇总后的事件类型，将其称为操作。这九种操作用于分割视频（图 3B1）。我们统计具有相同操作e 的日志，得到服务记录向量$E = [(e_1, n_1), (e_2, n_2) \ ...]$，其中$n_i$是连续$e_i$的计数。

4.2 操作锚点

操作锚点的主要目的是减轻分析人员观看视频时进行状态判定的负担（如第 2.3 节所述）。操作锚点可以对服务记录中的操作进行分段，并为各段赋予语义含义（**R4**）。它们还引入事件结构来概括视频内容。为优先处理异常事件（**R5**），我们采用基于相似度的方法查找时间异常，并采用马尔可夫技术检测服务记录中的序列异常

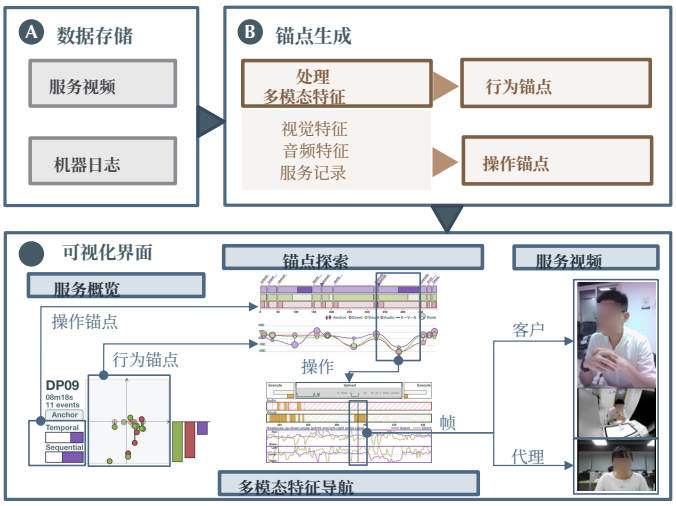


图 2。*Anchorage* 的系统架构包含三个模块：数据存储、锚点生成模块和可视化界面。

。随后将服务记录向量 E 输入以下算法，以获得相应的异常分数。

时间异常用于定位时长异常的操作。主成分分析（PCA）[62] 是一种广泛用于系统日志分析的无监督方法，可检测异常离散事件。它基于异常序列应与正常序列不同这一假设，计算输入序列与带标签序列之间的相似度。将标记为正常E_n的服务记录按操作进一步汇总，得到定长向量。随后将其降至k 个主成分，以构建正常空间S_n。若 $||y||^2 > Q_{1-\alpha}$，则认为该服务记录异常，其中y 是向S_n的投影长度，而$Q_{1-\alpha}$ 是置信阈值，默认为 95%。我们还考虑过系统日志分析中另一种常用方法——不变量挖掘 [63]。然而，该方法专为软件系统中的严格流程设计，泛化能力有限。与此同时，PCA具有可解释性强且不需要大型训练集的优点。

序列异常用于定位服务记录中不常见的时间顺序。马尔可夫链模型 [64]学习正常序列中各时间帧上不同离散状态的转移概率分布P。它为服务向量E 赋予：

$$P(E) = P(e_1, e_2, ..., e_T) = \prod_{t=1}^{T-1} p_{e_t, e_{t+1}} \tag{1}$$

其中T 是固定的窗口大小。引入非零常数 ε ，以防止某个子序列未出现在训练集时其概率为零。我们对事件记录进行放大和降采样，以创建固定大小的事件向量。我们将操作的异常阈值设置为 $|1/n|$，使得转移在训练集中至少发生一次；对于服务，则将阈值设置为能够涵盖95%方差的常数。选择马尔可夫链模型有三个原因：(1) 不需要大量训练数据；(2) 通过增大窗口大小来容纳大量日志记录，具有良好的可扩展性；(3) 当操作偏离标准流程时，能够识别出具体的操作（**R4**）。此外，客户服务具有预定义流程（代理指南），可作为正常训练集。尽管马尔可夫模型并非为识别缺失和过量事件而设计 [53], ，但其异常分数仍会反映这些情况

因为这些操作会出现在错误的位置。马尔可夫模型非常适合检测重复和不同步的操作（**R2**）。

4.3 行为锚点

行为锚点用于进行多模态满意度评估。与 [11], [24], 类似，我们采用线性模型生成客户满意度评分。我们扩展了该模型，使其涵盖事件持续时间，而不仅仅是情感状态（**R2**）。该模型结合面部表情 v、音频情绪 a和事件 e 来评估满意度。某项服务的客户满意度评分CS_s 计算如下：

$$CS_s = w_v f(\sum_{i=1}^{N} m_v v_i) + w_a f(\sum_{j=1}^{M} m_a a_j) - w_e \sum_{t=1}^{T} z_{e,t} \qquad (2)$$

其中，N、M和T 分别表示帧、话语和操作的总数。w 表示各通道的权重，默认采用等权重。我们还通过将求和范围限制在单个操作和模态内，获得了每种模态下该操作的满意度评分CS_e 。f 表示跨所有服务进行的标准标准化。z_e 表示事件持续时间的 z 分数。我们在标准化之前按操作对其进行分组，因为重复操作通常持续时间更短，会使计算结果产生混淆。对于情绪反应，我们为每种离散情绪赋予一个幅度权重m ，并采用先前研究 [11]提出的方案。一般而言，积极情绪的取值为 +1.0，中性情绪为0.0，而消极情绪通常趋近于-1.0。根据领域文献 [12] 以及与**E5-7**的讨论，我们将愤怒的权重略微调整为-1.2，将厌恶的权重调整为-1.0。较大的CS_s 值表示较高的满意度，反之亦然。上述所有设置均可重新配置，以适应其他需求。

5 可视化界面

Anchorage 的可视化界面支持多尺度满意度评估，以及通过直观的可视化设计进行锚点候选验证。图 3 展示了交互式系统的快照，并标注了（A）服务概览、（B）锚点探索视图、（C）多模态特征导航视图和（D）服务视频视图。服务视频视图展示原始服务视频，并在选择相应视频或事件时与其他视图同步播放。该视图还支持常规的视频播放功能，以及后续章节所述的其他交互操作。

5.1 服务概览和浮标图表

服务概览（图 3A）提供所有服务视频的概览。它列出所有视频，并支持在多个视频之间进行快速比较，以查找感兴趣的服务。每个列表项（图 3A1）包含三列，用于显示不同的满意度指标（**R1**）。可视化界面的颜色编码保持统一（视觉使用绿色，音频使用红色，事件使用紫色）。最左侧列显示视频的基本信息。水平条形图展示第 4.2 节所述的时间异常分数和序列异常分数。已识别的异常以实色表示，正常服务则使用条纹色表示。最右侧列为垂直条形图，展示第 4.3 节所述不同模态的满意度分数CS_s 。

中间一列是基于散点的设计，称为浮标图 （图 4C）。它将各项操作的多模态满意度分数汇总到一张图表中。每项操作由两个圆点表示，圆点以不同颜色标示其所属类型（视觉 或音频）。圆点的垂直位置表示该类型的满意度评分CS_e 。靠近水平线的圆点不透明度较低，因此在视觉上不太显眼。所有圆点的水平位置表示反向事件评分。反向设计旨在符合象限启发式：将表示时长较短的负分数转换为正数。此外，同一项操作的两个圆点会通过垂直连线相连，依据连续性原则呈现一个连通视觉组件。垂直和水平刻度均以零为中心，并限制在阈值范围内。

设计依据。 我们设计了一种隐喻，用于降低浮标图的学习曲线，降低可视化素养的门槛 [65]。浮标图采用浮标隐喻，因为浮标会连接到锚上。一个点被称为浮标。正常服务应包含许多常规操作；因此，各点聚集在零附近。虽然这些彼此接近的浮标漂浮在水面上，几乎难以察觉，但显著的浮标会下沉并上升，成为视觉上明显的锚点。这些离群值凸显了具有潜在满意度模式的异常操作（**R5**）。当极化的锚点与接近零的聚合分数同时出现时（例如，图 3A1 的音频通道），它们能够检测出抵消情形（**R3**）。

此外，浮标图还可以结合图 4A 中的象限启发式进行解读。我们通过检查包含最多锚点的象限来估计总体满意度。例如，图 3A3 显示了许多漂浮的浮标和少量下沉的锚点。锚点的位置表明，客户在少数持续时间较长的操作中表现出负面情绪。这展示了浮标图如何高效地总结区分性模式（**R1**）。浮标图是多模态融合的视觉替代方案（例如，公式 2），而多模态融合可能需要额外工作来优化代价函数。

浮标图能够有效编码多模态特征。图 4C 展示了Wang等人 [35] (e_{1-3} 主导型、e_4 互补型，以及e_5 冲突模态）的三种模态间交互类型所呈现的视觉模式。我们以图 3 中的四个案例为例：A2-A3 在一个通道中主要表达强烈情绪；A4 通过互补两个通道传达负面情绪；A1 则在某些操作中表现出冲突行为。这些视觉线索有助于分析客户的情绪特征（**R2**），为后续分析提供依据。

浮标图可以扩展，以应对各种设计问题。适用于基于散点的设计的技术也可能适用于浮标图。针对可扩展性问题，我们可以减小点大小，并在两侧叠加条形图，以观察人口密集区域中的操作分布。为了优先关注最重要的项目，我们可以使用基于象限的热力图，按照感兴趣的锚点模式筛选并排列视频。

设计备选方案。 我们专注于设计直观、标准化的图表，以适应目标用户——服务提供者和代理——背景各异的特点。构建抽象表示是处理大量视频集合的常用策略。我们采用降维技术生成视频簇，并将离群值可视化，以减少审阅工作量。然而，这些技术忽视了时间关系，也无法检测相互抵消案例。尽管我们可以设置示例来引导聚类，但要让新手了解技术假设并避免过度依赖无监督结果，仍然颇具挑战。因此，我们采用了“依次查看”策略 [66]，并将不同模态下的满意度分数作为排序顺序。

作为浮标图表的替代方案，我们考虑使用堆叠条形图按顺序展示操作，如图4D所示。当多模态分数占主导时，视觉上很容易看出来

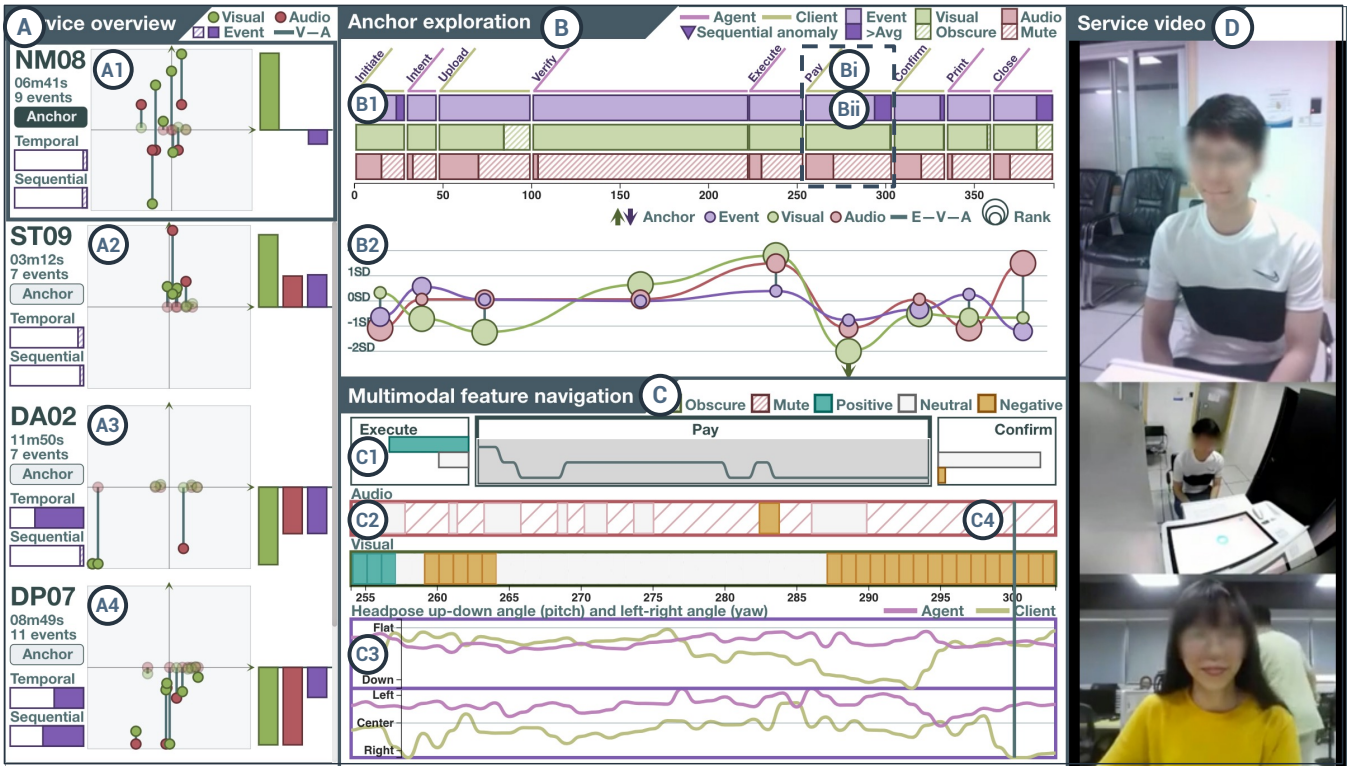


图3. *Anchorage* 通过提高顺序分辨率和时间分辨率，放大呈现客户满意度指标。服务概览（A）根据客户的响应和服务流程的流畅度，展示初始满意度评分。锚点探索视图（B）展示按操作分段的服务中的客户满意度变化，并突出显示异常。多模态特征导航视图（C）提供详细的多模态信息，用于验证满意度分数，并支持交互式导航至服务视频视图（D）。

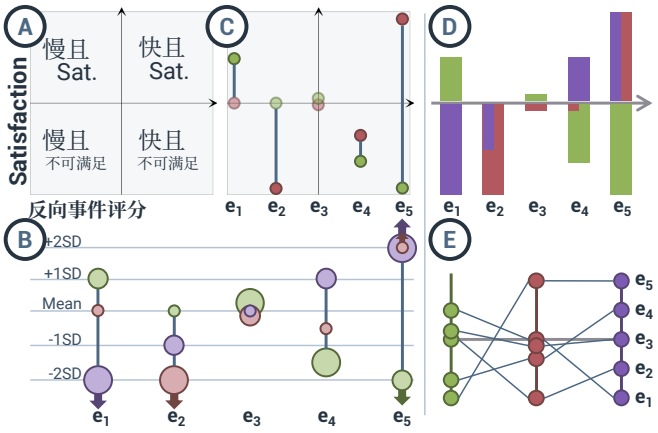


图4. *Anchorage* 采用浮标隐喻，结合流程理解来概括多模态满意度评估。（A）展示了浮标图的四象限启发式方法。（B）为横向浮标图，逐步绘制五项操作的多模态满意度分数（e_{1-5}）。（C）为浮标图，在单张图表中概括一项服务的所有操作。（D-E）为（C）的替代设计。

。然而，互补型和相互冲突的模态由于缺乏快速决策规则，降低了图表的可解释性。图表同时显示积极值和消极值时也会造成混淆，而且随着操作数量增加，还会出现可扩展性问题。另一个选项是平行坐标（图4E）。然而，较强的模态内可扩展性无法弥补进行跨模态比较时线条造成的视觉杂乱。比较三个同时分析多种模态以查找异常操作。图 4C-E 使用相同的数据集，以突出显示它们之间的差异。

5.2 锚点探索视图

此视图（图 3B）支持基于服务记录的操作级锚点探索。基于时间线的可视化（图 3B1）用一列视觉组件表示每项操作。水平位置表示服务时间。每列包含四行。在顶行（图 3Bi）中，我们可视化操作、轮流发言信息和序列异常指示器，以提供流程上下文并指出序列不一致（**R4**）。代理发言时，线条显示为粉色；客户发言时，则显示为黄色。三角形图标表示该操作存在序列异常，如图 5A 所示。

底部的三条柱（图 3Bii）分别展示事件、视觉和音频模态的统计信息。第一条紫色柱表示操作的时长。深紫色部分表示超出操作平均时长的时间。这表明该操作比平时耗时更长，可视为时间异常。第二条绿色柱和第三条红色柱分别表示视觉和音频通道中检测到特征的时间比例。条纹图案表示缺少特征，其中绿色代表客户面部被遮挡，红色代表静音。例如，图 3Bii 显示，整个操作过程中客户面部都未被遮挡，而对话时长约占总时长的三分之一。这些指标提供了背景信息，有助于了解检测到的特征的可靠性。视觉组件与其他视图相关联，因此点击它们可以跳转至多模态特征和原始服务视频（**R6**）。

横向浮标图（图 3B2）展示满意度变化。图表的水平编码沿用上方的时间线，因此所有点都位于操作时段的中间位置。图 4B-C 展示了该图与浮标图之间的对应关系。虽然两种图表采用相同的隐喻，但仍有细微差别。每项操作由三个点表示，其中包括事件。在这里，垂直位置采用 z 分数来统一所有模态。视觉和音频分数在整个操作期间求和，并在选定服务内进一步标准化。例如，e_5 在图 4B 中的所有模态分数都明显偏离，而e_3 对应的分数则处于平均水平。锚图标表示数值高于两个标准差。该刻度有助于检测最异常的服务操作（**R5**）。浮标的大小编码绝对偏差排名，以突出最具影响力的锚。偏差越显著，浮标越大。绘制顺序优先绘制较小的浮标，以避免遮挡和视觉杂乱（见图 5A）。

设计依据。横向浮标图弥合了浮标图与时间线之间的差距。我们没有在时间线上明确编码短于平均值的操作，因为这可能会扭曲布局。此外，这些操作的重要性较低，因为较短的事件通常需要验证的行为锚点较少。通过引入事件浮标，横向浮标图弥补了缺失的时间异常信息。采用熟悉的视觉元素和隐喻，可以减轻学习新型可视化的负担。当用户熟悉该系统时，这种对应关系能够提升效率。

设计备选方案。 在设计过程中，我们参考了事件序列设计空间 [50] ，并迅速排除了基于层次结构、桑基图和基于矩阵的设计，因为它们不适合相关任务。为了简洁且易于理解，我们将条形图与时间线结合，构建了当前设计。在可视化特征变化方面，Zeng 等人 [25] 提出了五种用于展示情绪流的设计。然而，由于缺少事件上下文，这些设计的可视化摘要能力有限。我们的横向浮标图结合了多种视觉元素，以连贯地呈现服务操作中的满意度变化。

5.3 多模态特征导航视图

此视图支持对原始视频进行交互式导航（**R6**）。我们采用外围图 [67] 作为操作摘要（图 3C1）。在中间的聚焦细节视图中，我们融合面部和音频特征，为i帧分配一个激活值$v_i = \{-1, 0, 1\}$。融合时优先考虑非中性情绪，其中消极情绪的优先级更高，因为它们对满意度的影响更大 [12]。我们绘制激活值以展示操作概览。刷选可选取下方特征对应的时间段。两侧的外围图便于快速导航至连续操作，并通过相邻操作为聚焦的操作提供上下文。三条柱状图显示激活值的数量。

音频和视觉通道是用于编码面部表情和音频情绪的积极与消极输出的一维阴影矩阵（图 3C2）。遮挡和静音的帧与之前一样，使用条纹图案进行编码。头部姿态信息以折线图显示（图 3C3）。图表刻度直接标注了角度的语义含义。视觉特征与头部姿态相邻，用户因此无需低头查看，即可验证存在遮挡时的情绪。点击这些视图（图 3C4）即可跳转到视频中的相应时间点。

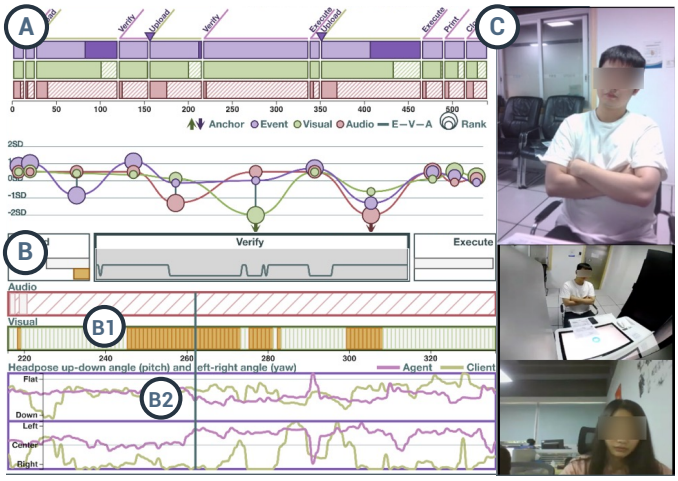


图 5. 不满意案例显示，客户因反复的操作和注意力不集中的服务人员而感到恼火。

6 评估

我们通过与 **E5** 开展案例研究，并进行结构化用户研究，展示*Anchorage*的有效性和可用性。

6.1 案例研究

本研究介绍了**E5** 对其两名实际下属**S1** 和**S2**所提供服务的满意度评估过程。他的任务是通过查看按**S1** 和**S2**筛选的临时构建的数据集（第 3.3 节），对服务和服务人员进行评分。他密切参与了问题特征描述和临时构建的数据集的构建，但在此次案例研究之前，他既未使用过*Anchorage* ，也未看过这些视频。他了解一线服务人员通常熟知的满意度场景。研究鼓励他在探索过程中遵循出声思维协议。

服务探索（R1-2）。从服务概览（图 3A）开始，**E5** 首先按照服务满意度得分降序排列视频。ST09（图 3A2；探索过程中视频名称已被隐藏）在列表中排名第一。他注意到所有指标都倾向于该服务，因为所有条形图都显示积极值，并且浮标图中有一个明显的音频锚点。阅读左栏中的基本信息后，他很快确定这应属于**ST** 类型。他从服务视频视图中选择该服务，以查看代理是谁（**S1**）。随后，他继续浏览其他视频。尝试了几次后，NM08（图 3A1）引起了他的注意。他查看了接近于零的音频满意度得分以及浮标图中两极分化的浮标，怀疑这是一个相互抵消案例。他对这个案例产生了兴趣，想了解其中的冲突行为。他选择该服务，继续探索服务记录。

操作探索（R3-5）。查看锚点探索视图（图3B）时，他首先检查了流程摘要（图3B1），没有发现多少顺序或时间异常。代理的大多数操作都及时完成，且没有流程偏离代理指南。他排除了**DP** 类型。随后，他查看横向浮标图（图 3B2），发现了表明面部表情明显消极的视觉锚点。他还注意到，该锚点前一次操作时的表现非常积极。重新查看操作名称（“执行”和“支付”）后，他对事件有了一些头绪，但还需要更多证据。他点击锚点图标，调查关键转折时刻。

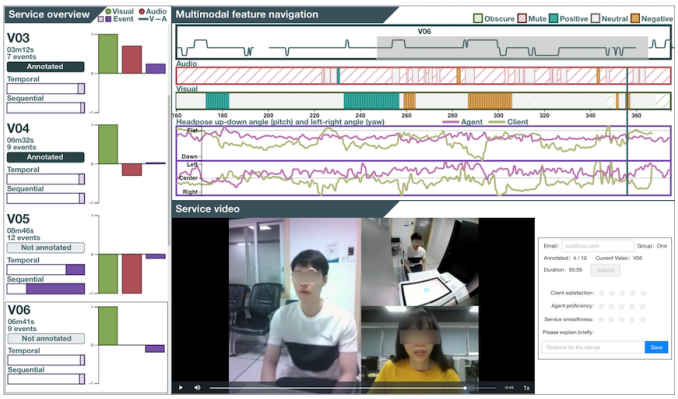


图 6. 用户研究中使用的基线系统界面。该界面将客户满意度评分和多模态特征可视化，以提供基本的计算支持。该界面由 *Anchorage* 修改而来，移除了基于事件的视觉组件，以分离它们的影响。

特征验证（R5-6）。进入多模态特征导航视图（图3C）后，他从外围摘要和视觉特征（图3C4）中发现，最明显的消极情绪出现在“支付”操作的后半段。他点击C4中的橙色帧，跳转到服务视频（图3D）。通过观看原始视频，他得出结论：负面情绪源于不得不支付，而非代理疏忽。他排除了**DA** 类型，并将其判定为**NM** 类型。然而，阅读左侧外围图（图3C1）时，他想知道为什么客户在前一次操作期间有许多积极行为。他点击该图并再次进行特征验证分析。他惊讶地发现，客户全程只是在手机上发短信。尽管视觉通道上的满意度评分很高，这进一步坚定了他对**NM**的判定。他确信*Anchorage*能够检测出相互抵消案例。**S1** 积累了若干**NM** 和**ST**案例，足以被评为表现良好。

满意度评估（R1-6）。通过按评分从低到高排列视频，**E5** 发现，无论在视觉还是音频通道中，DP07（图 3A4）的满意度评分都是最低的。他注意到，系统检测到该服务存在时间异常和序列异常。他在锚点探索视图中查看了这段服务（图5A）。他注意到，客户上传文件所用的时间比平时更长。实验流程还重复了两次，因此可能会惹恼客户。他断定这项服务属于 **对流程不满意**类型，因为他知道这是唯一符合的场景。

一个 DAP案例。“验证”操作下的视觉锚点吸引了**E5**的注意。他感到困惑，因为典型的**对流程不满意**场景并非如此。他调查了这个视觉锚点，并看到了图 5B1 中持续出现的橙色帧。从头部姿态图中，他观察到**S2**一直低着头。他点击了该行为（图5B2），并从视频中判断出**S2**当时正在玩手机（图 5C）。他推测**S2** 可能结合了**对服务人员不满意**和**对流程不满意**两种类型，形成了一个更加不满意的**DAP** 案例：冗长的实验流程和注意力不集中的服务人员令客户感到恼火。尽管如此，他仍认为**S2** 作为代理表现不佳，但作为业务分析师表现出色，因为她主动提出了新需求。

6.2 用户研究

我们采用被试间设计，并设置了两种系统条件，以评估 *Anchorage* 在评估客户服务视频方面的有效性和可用性。

实验装置。 在可视分析系统中呈现多模态特征，通常比缺乏计算支持的基线系统带来更好的任务表现 [32], [33], [44]。然而，由于交互差距较大且存在复合效应，将同时包含操作锚点和行为锚点的*Anchorage* 与一个简单的基线系统进行比较可能并不公平。本研究旨在评估引入事件上下文（即操作锚点）是否能提升满意度评估任务的表现。我们通过移除*Anchorage*中基于事件的可视化组件——即浮标图、锚点探索视图和外围图——创建了基线系统（图 6）。其余部分则用于可视化有助于自动化满意度评估的多模态特征 [9], [11]。我们记录了用户的鼠标操作，以进行溯源分析。

数据与任务。我们从第3.3节所述的数据集中抽取视频，用于满意度评估任务。根据一项包含**E5-6**的先导研究，我们估计参与者可以在十分钟内标注三个视频。考虑到工作量和时长，我们随机选取了12个视频用于研究。这些视频的选取遵循两个条件：1）涵盖所有四种满意度类型，且每种类型数量相同；2）由不同客户出演。这些条件确保样本具有多样性和独立性。我们使用两个视频进行演示，参与者则评估其余视频。剩下的十个视频总时长为68.6分钟，构成十项满意度评估任务（**图7中的T1-T10**，视频名称已做掩码处理）。除了要求参与者使用从1（低）到5（高）的李克特量表评价客户满意度外，我们还请他们评估客服人员熟练度和服务流畅度，以便进一步区分这些视频。

参与者。 我们采用滚雪球抽样，从**E5-7** 的同事开始，在其所在公司的合作者中招募了24名参与者（女性8名，男性16名；年龄：均值（M） $=28.4$，标准差（SD） $=5.3$）。其中16名参与者是科学、技术、工程和数学学科的本科生，其余参与者则拥有不同领域的文凭。他们在客户服务方面拥有1至13年的相关经验（$M = 5.1$, $SD = 3.7$）。完成研究后，他们获得了等值于7.50美元的报酬。我们随机分配参与者使用*Anchorage* 系统（P_A）或基线 系统（P_B）。

实验流程。 由于隔离限制，本研究以远程方式开展。参与者可以访问部署在网上的指定系统。我们首先征得他们的知情同意，并通过录制的视频介绍研究背景和系统功能，时长约为 12 分钟。参与者使用训练示例自由探索三分钟后，应使用指定系统完成十项任务。由于我们没有对任务设置时限，因此提供现金奖励，以避免参与者提交低质量的回答。如果参与者给出的每项满意度评分与**E5-6** 的评分相差不超过一个分值，则可获得 1.5 美元（最多可获五次奖励）。最后，他们填写了一份关于指定系统及其背景信息的问卷。所有场次持续时间为 30–90 分钟（$M = 56.5$, $SD = 18.3$）。

6.2.1 结果

Anchorage 组的完成时间短于基线组，而标注基本保持不变。

完成时间。我们比较了 Anchorage组（A）和基线组（B）完成十项评估任务所用的时间（分钟）。在评估服务视频时，使用 Anchorage（$M_A = 27.6, SD_A = 19.3$）平均比使用基线（$M_B = 35.3, SD_B = 17.1$）快 21.8%，但曼-惠特尼 U 检验表明，这一差异在统计学上并不显著（$W = 51.0, p = 0.118$）。有两项单独任务

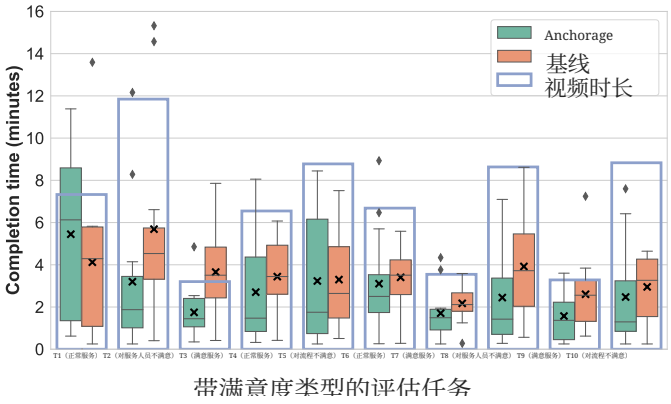


图7. 各项任务的完成时间及对应视频的时长。 ×表示平均完成时间。

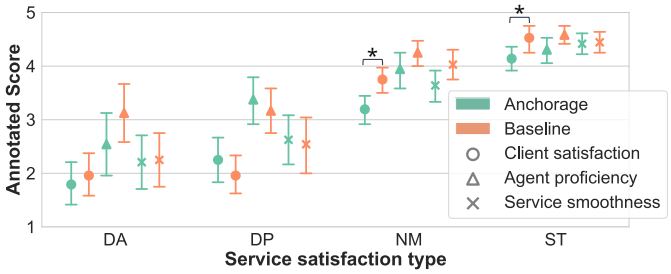


图8. 各视频的平均标注评分及其95%置信区间。∗ 表示统计显著差异。

对 Anchorage组而言，完成速度显著更快，分别是 T2（$U = 38, p < 0.05$）和 T3（$U = 30, p < 0.01$）。结合案例研究中的发现，在事件上下文中，Anchorage 用户可能更容易对两种满意度类型（**对服务人员不满意** 和 **满意服务**）作出初步决策。从图7中我们观察到，除 T1 外，Anchorage组在所有任务中的平均完成时间和完成时间中位数都更短。由于 T1 是第一项任务，一个可能的原因是，学习使用 Anchorage系统所需的时间比学习基线所需的时间更长。

Anchorage组的完成时间方差较大。九位P_A 完成十项任务的速度快于平均水平（即， ＜27.6 分钟），而只有三位P_B 在该时间内完成。通过分析三位P_A （他们需要超过50分钟）的分析溯源，我们发现，他们倾向于观看原始服务视频，而不是利用锚点来确定调查工作的优先顺序。这种情况只出现在一位P_B身上。相反，观察发现，十分钟内完成任务的P_A专注于验证锚点并核实特征。我们预计，用户熟悉该系统后，这将成为常态。

标注评分。我们比较了 Anchorage 与基线组中每种满意度类型的三项评估指标。图8展示了不同满意度类型在所有指标上的具体差异。结果表明，只有两组在**正常服务**（$U = 32.5$，$p < 0.05$）和**满意服务**（$U = 39.5$，$p < 0.05$）的客户满意度评分差异具有统计显著性。总体而言，P_B 倾向于给**正常服务**和**满意服务**更高的满意度评分，而P_A的评分则更接近**E5-6**提供的标签。

我们进一步研究了这两个系统能否清楚地区分不同的满意度类型。我们进行了 Wilcoxon符号秩检验，以适应任务之间可能存在的依赖关系——

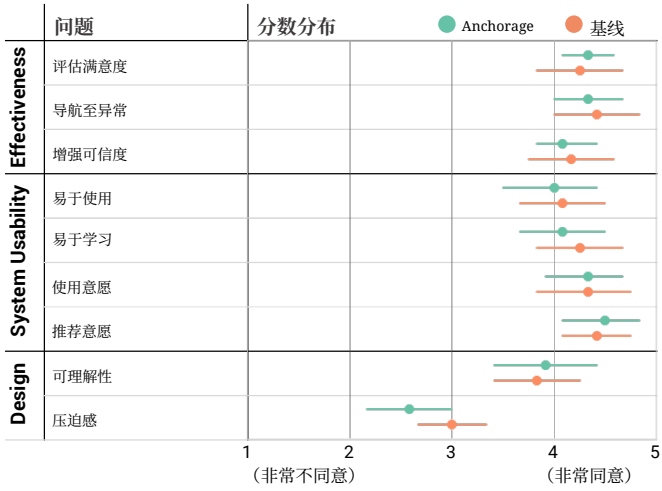


图9. 两组参与者在系统整体有效性、系统可用性和视觉设计方面的问卷结果。

。对于两个组，均无证据表明**NM**与**ST** 之间的客服人员熟练度存在显著差异（$W_A = 15$，$p_A = 0.202$；$W_B = 5$，$p_B = 0.068$）。Anchorage组中满意度类型之间的其他所有两两比较均有$p < 0.05$，这意味着P_A 能够清楚地区分这四种满意度类型。然而，对于基线组，**DA** 与**DP**这组比较中，客户满意度（$W = 10.5$，$p = 1$）、客服人员熟练度（$W = 18$，$p = 1$）和服务流畅度（$W = 7.5$，$p = 0.262$）均无显著差异。这表明，在客户服务中缺少事件上下文的支持时，用户可能会认为**DA**和**DP** 类型相同。这种标注方式可能对并未造成不满意案例的代理有失公允。因此，在满意度评估中应考虑事件上下文，即使采用自动化方法也是如此，以避免对代理产生偏差。

问卷。图9报告了两组在三个维度上的问卷结果对比：任务有效性、系统可用性和视觉设计。图9所列的所有指标在两个系统之间均未发现统计显著性。从业者欢迎使用多模态行为特征来支持满意度评估任务。系统还能够帮助用户快速定位异常。与仅使用多模态行为锚点的基线系统相比，使用带有操作锚点的 Anchorage 时，用户能够区分出更多异常（**对服务人员不满意**与**对流程不满意**）。借助可视分析系统，用户对自动化结果更有信心。一些P_A 评论说，单项操作的评分比服务评分更准确。这证实了我们的假设：引入事件结构有助于评估满意度。

参与者总体上对这些系统的系统可用性和视觉设计给予了积极反馈。一个有趣的发现是，尽管基线系统更容易学习，但其可理解性和压迫感似乎不如*Anchorage*。学习曲线上的细微差异也可能体现在 T1 较慢的完成时间上。Anchorage 采用了更多视觉组件和新颖的可视化设计，因此比基线的压迫感更低。这可能归因于参与者对事件上下文的熟悉，以及隐喻式设计的直观性。Anchorage组的参与者报告了更多与事件相关的洞见，使用了象限启发式，并采用了锚点类比作为评分依据。

7 讨论

我们总结了在客户服务场景中开展满意度分析设计研究期间获得的经验教训如下。

客户服务的可迁移性。我们采用公共服务来刻画问题领域。抽象出的工作流程可迁移至在线购物中的客户服务对话数据集 [54], , 如图1所示。这是因为客户服务具有目标导向型任务和协作式沟通的特点。由于我们的服务类型侧重于处理申请，因此与联络中心通话相比，我们的数据中的对话更加稀疏 [8], [9]。不过，我们也捕捉了客户面部表情等密集特征，并设计了同时支持这两类特征的系统。这增强了该方法向其他服务类型的可迁移性，无论其特征是否稀疏。

通过事件分析构建视频结构。由于难以理解视频中的状态并检测事件，视频通常被归类为非结构化数据。在视频以序列记录形式录制的情况下，我们可以采用事件分析来制定细粒度的视频分割方案。随着远程工作环境日益普及，这类场景将越来越常见。本文所述框架可推广至其他应用，例如在线教育、智能制造、系统界面测试以及虚拟现实中的交互。我们展示了在传统的基于视频的满意度分析流程中引入操作锚点，能够简化对视频内容的理解。视频分析应积极利用事件结构，为从视频中提取的特征提供框架。

临时构建的数据集作为 VAST 的测试案例。收集真实数据集是 VAST 系统（包括视频分析）面临的最艰巨挑战之一 [55], 。在满意度分析中，极端案例（非常满意和非常不满意）在现实生活中十分罕见。我们使用临时构建的数据集来应对样本不平衡这一挑战。通过与领域专家密切合作，我们定义了典型的满意度模式，并通过模拟收集这些模式。它们提供的真实标注可用于预训练模型，并作为初始参考点，以缓解冷启动问题。我们还可以控制环境，隔离不必要的影响，使结果具有可比性。

让领域专家参与创建此类数据集对我们的研究很有帮助。该数据集为合作者和我们提供了共同基础，使我们能够在设计过程中讨论对视觉设计的期望。然而，在我们的案例中，由于采用了模拟方式，客户回应的真实性和动态性值得商榷。例如，顾客感到流程漫长乏味时，可能会更频繁地看手机 [15], , 而我们却要求受试者更加专注。我们还应注意观察者悖论，即实验者的在场会影响数据收集 [17], [18]。应进一步讨论并制定更标准化的规范，以促进公平、公正的评估。

隐私与多模态特征。在客户服务中，录制服务过程是一种常见做法，可防止因错误信息而引发冲突 [10]。当需要调查投诉时，视频尤其有用。本研究收集客户服务视频时，已获得内部机构审查委员会批准（#HREP- 2021-0162）及参与者的书面知情同意。然而，日益严格的隐私政策可能会禁止收集包含客户面部等敏感特征的服务视频。由于 *Anchorage*利用了从视频中提取的许多情绪特征如果禁止收集视频，其有效性将受到显著影响。

然而，我们已经证明，结合操作锚点和行为锚点可以提升满意度评估的表现，同时不影响标注精度。还有更多值得探索的多模态特征，或许能够促进满意度评估。例如，事件特征（例如业务流程）以及其他行为特征（例如机器操作和客服人员行为）为满意度的因果分析提供了背景信息。屏幕空间和数据空间脱敏技术 [68] 或许有助于保护客户隐私。我们还收集了关于客户的多视角视频，可提供额外视角以推断他们的行为。未来的研究可以探索用于总结多视角视频的可视化技术。另一个尚未得到充分探索的特征是客服人员的情绪。客服人员与客户之间的情感互动可以帮助推断对方的情感状态，并评估客服人员是否做出了恰当的反应。

8 结论

本文探讨了如何利用客户服务视频和服务记录评估客户满意度。从视频中提取的多模态行为特征经过融合，可用于初步评估满意度。我们引入机器日志，以实现语义上有意义的视频理解，并通过更高的序列和时间分辨率丰富传统的满意度评分。这两者共同构成锚点概念。我们利用多视角异常检测框架构建锚点，以缩小广阔的事件空间。我们开发了浮标图和多维视图，以有效总结服务内容，并引导用户浏览感兴趣的片段。

我们构建了一个临时数据集，以展示*Anchorage*能够检测令人满意、不满意以及中和的正常案例。视频分析与事件序列分析相结合，在有效理解视频方面展现出良好的前景。我们发现，将事件上下文引入视频分析，可以提升视频中客户满意度评估的表现。我们的方法可用于收集无标注非结构化视频及其序列记录的场景。

致谢

我们感谢 Wucong Chen 和 Guang Fo Zhan Qu 团队在用户研究中提供的宝贵反馈与协调工作。我们也感谢匿名评审提出的建设性意见。本研究部分由资助项目 FSUST19-CWB09 支持。

参考文献

[1] R. L. Oliver，《满意度：消费者的行为视角》，第2版。Routledge，2010年。[2] “质量管理——客户满意度——监视和测量指南”，ISO 10004:2018，2018年。[3] A. Wong，“服务接触中的情感满意度作用”，《服务质量管理：国际期刊》，第14卷，第5期，第365–376页，2004年。[4] B. Tronvoll，“负面情绪及其对客户投诉行为的影响”，《服务管理期刊》，第22卷，第134页，2011年，第1期，第111页。[5] R. A. Peterson 和 W. R. Wilson，“测量客户满意度：事实与伪象”，《营销科学学报》，第20卷，第1期，第61–71页，1992年。[6] S. Al-Otaibi、A. Alnassar、A. Alshahrani、A. Al-Mubarak、S. Albugami、N. Almutiri 和 A. Albugami，“使用情感分析测量客户满意度”，《国际高级计算机科学与应用期刊》，第9卷，第2期，第106–117页，2018年。

[7] A. See 和 C. Manning，“理解和预测神经生成式聊天机器人中的用户不满意度”，载于 SIGDIAL 会议论文集。新加坡及线上：计算语言学协会，2021年7月，第1–12页。[8] Y. Park 和 S. C. Gates，“利用自动生成的通话转录文本实现客户满意度的实时测量”，载于 CIKM 会议论文集。纽约：ACM，2009年，第1387–1396页。[9] A. Ando、R. Masumura、H. Kamiyama、S. Kobashikawa、Y. Aono 和 T. Toda，“基于分层多任务模型的联络中心通话客户满意度估计”，IEEE/ACM 音频、语音与语言处理汇刊，第28卷，第715–728页，2020年。[10] M. Saberi、O. K. Hussain 和 E. Chang，“联络中心的过去、现在与未来：文献综述”，《业务流程管理杂志》，2017年。[11] K. P. Seng 和 L.-M. Ang，“联络中心的客户情绪与满意度视频分析”，IEEE 人机系统汇刊，第48卷，第3期，第266–278页，2018年。[12] V. Liljander 和 T. Strandvik，“服务满意度中的情绪”，《服务行业管理国际期刊》，第8卷，第169页，1997年，第2期，第148页。[13] A. Cheshin、A. Amit 和 G. A. van Kleef，“客户服务中情绪强度的人际影响：感知适当性和真实性如何塑造服务人员的情绪表达、客户信任与满意度”，《组织行为与人类决策过程》，第144卷，第97–111页，2018年。[14] Q. Zhang、W. Wang 和 Y. Chen，“前沿：利用逐时刻非结构化数据进行消费过程中的社交聆听：电影欣赏与实时评论案例”，《营销科学》，第39卷，第2期，第285–295页，2020年。[15] Z. X. Liu、Y. Liu 和 X. Gao，“利用移动眼动追踪评估对服务办公室的满意度”，载于《设计、用户体验与可用性：实践与案例研究》。查姆：施普林格国际出版社，2019年，第183–195页。[16] S. Kumar、M. Yadava 和 P. P. Roy，“融合脑电图反应与产品评论情感分析以预测客户满意度”，《信息融合》，第52卷，第41–52页，2019年。[17] H. Gunes 和 B. Schuller，“连续输入中的分类与维度情感分析：当前趋势与未来方向”，《图像与视觉计算》，第31卷，第2期，第120–136页，2013年。[18] D. McDuff、R. E. Kaliouby 和 R. W. Picard，“众包收集对在线视频的面部反应”，IEEE 情感计算汇刊，第3卷，第4期，第456–468页，2012年。[19] A. Generosi、S. Ceccacci 和 M. Mengoni，“用于追踪和分析零售店客户行为的深度学习系统”，载于 ICCE-Berlin 会议论文集，2018年，第1–6页。[20] D. McDuff、R. E. Kaliouby、J. F. Cohn 和 R. W. Picard，“预测广告喜好度与购买意向：对广告面部反应的大规模分析”，IEEE 情感计算汇刊，第6卷，第235页，2015年，第3期，第223页。[21] M. Slim、R. Kachouri 和 A. B. Atitallah，“基于最显著面部情绪的客户满意度测量”，载于 SSD 会议论文集，2018年，第502–507页。[22] N. Sugianto、D. Tjondronegoro 和 B. Tydd，“利用视频监控进行客户满意度分析的深度残差学习”，载于 AVSS 会议论文集，2018 年，第 1–6 页。[23] M. Gonz´alez-Rodr´ıguez、M. D´ıaz-Fern´andez 和 C. Pacheco G´omez，“面部表情识别：通过情绪测量游客满意度的新兴方法”，《远程信息处理与信息学》，第51卷，第101404页，2020年。[24] G. Yolcu、I. Oztel、S. Kazan、C. Oz 和 F. Bunyak，“用于监测客户兴趣的深度学习面部分析系统”，《环境智能与人性化计算杂志》，第11卷，第1期，第237–248页，2020年。[25] H. Zeng、X. Wang、A. Wu、Y. Wang、Q. Li、A. Endert 和 H. Qu，“EmoCo：演示视频中情绪连贯性的可视分析”，IEEE 可视化与计算机图形学汇刊，第26卷，第1期，第927–937页，2020年。[26] C.-X. Ma、J.-C. Song、Q. Zhu、K. Maher、Z.-Y. Huang 和 H.-A. Wang，“EmotionMap：地图上的视频情绪内容可视分析”，《计算机科学与技术杂志》，第35卷，第591页，2020年，第3期，第576页。[27] X. Wang、H. Zeng、Y. Wang、A. Wu、Z. Sun、X. Ma 和 H. Qu，“Voicecoach：面向公开演讲声音调节技巧的交互式循证训练”，载于 CHI 会议论文集。美国纽约州纽约市：ACM，2020年，第1–12页。[28] X. Wang、Y. Ming、T. Wu、H. Zeng、Y. Wang 和 H. Qu，“Dehumor：幽默分解的可视分析”，IEEE 可视化与计算机图形学汇刊，第28卷，第12期，第4609–4623页，2022年。

[29] K. Maher、Z. Huang、J. Song、X. Deng、Y.-K. Lai、C. Ma、H. Wang、Y.-J. Liu 和 H. Wang，《E-ffective：一种用于探索励志演讲情绪与有效性的可视分析系统》，IEEE 可视化与计算机图形学汇刊，第 28 卷，517，2022 年 1，508，第 – 页。[30] H. Zeng、X. Wang、Y. Wang、A. Wu、T.-C. Pong 和 H. Qu，《Gesturelens：演示视频中手势的可视分析》，IEEE 可视化与计算机图形学汇刊，第 1–1 页，2022年。[31] H. Zeng、X. Shu、Y. Wang、Y. Wang、L. Zhang、T.-C. Pong 和 H. Qu，《EmotionCues：课堂视频的面向情绪的可视化摘要》，IEEE 可视化与计算机图形学汇刊，第 27 卷，第 7 期，第 3168–3181 页，2021年。[32] H. Li、M. Xu、Y. Wang、H. Wei 和 H. Qu，《一种用于辅助在线考试监考的可视分析方法》，载于 CHI 会议论文集。纽约：ACM，2021年。[33] T. Tang、Y. Wu、Y. Wu、L. Yu 和 Y. Li，《Videomoderator：一种面向电子商务多模态视频审核的风险感知框架》，IEEE 可视化与计算机图形学汇刊，第 28 卷，第 1 期，第 846–856 页，2022年。[34] P. P. Liang、Y. Lyu、G. Chhablani、N. Jain、Z. Deng、X. Wang、L.- P. Morency 和 R. Salakhutdinov，《Multiviz：用于多模态模型可视化与理解的分析基准》，ArXiv 预印本 ArXiv:2207.00056，2022年。[35] X. Wang、J. He、Z. Jin、M. Yang、Y. Wang 和 H. Qu，《M2Lens：用于情感分析的多模态模型可视化与解释》，IEEE 可视化与计算机图形学汇刊，第 28 卷，第 1 期，第 802–812 页，2022年。[36] A. Wu 和 H. Qu，《视频集合的多模态分析：TED演讲中演示技巧的可视探索》，IEEE 可视化与计算机图形学汇刊，第 26 卷，2442，2020 年 7，2429，第 – 页。[37] T. Blascheck、F. Beck、S. Baltes、T. Ertl 和 D. Weiskopf，《富数据用户行为的可视分析与编码》，载于 VAST 会议论文集，2016年，第 141–150 页。[38] E. J. Soure、E. Kuang、M. Fan 和 J. Zhao，《Coux：针对数字界面的出声思维可用性测试视频进行协作式可视分析》，IEEE 可视化与计算机图形学汇刊，第 28 卷，第 1 期，第 643–653 页，2022年。[39] Y. Lin、K. Wong、Y. Wang、R. Zhang、B. Dong、H. Qu 和 Q. Zheng，《Taxthemis：可疑逃税团伙的交互式挖掘与探索》，IEEE 可视化与计算机图形学汇刊，第 27 卷，第 2 期，第 849–859 页，2021年。[40] W. Zhang、J. K. Wong、X. Wang、Y. Gong、R. Zhu、K. Liu、Z. Yan、S. Tan、H. Qu、S. Chen 和 W. Chen，《Cohortva：一种基于历史数据交互式探索队列的可视分析系统》，IEEE 可视化与计算机图形学汇刊，第 29 卷，第 1 期，第 756–766 页，2023年。[41] B. H¨oferlin、M. H¨oferlin、G. Heidemann 和 D. Weiskopf，《可扩展视频可视分析》，Information Visualization，第 14 卷，第 1 期，第 10–26 页，2015年。[42] K. Kurzhals、M. John、F. Heimerl、P. Kuznecov 和 D. Weiskopf，《电影可视分析》，IEEE多媒体汇刊，第 18 卷，第 11 期，第 2149–2160 页，2016年。[43] Z. Chen、S. Ye、X. Chu、H. Xia、H. Zhang、H. Qu 和 Y. Wu，《使用 viscommentator 增强体育视频》，IEEE 可视化与计算机图形学汇刊，第 28 卷，第 1 期，第 824–834 页，2022年。[44] D. Deng、J. Wu、J. Wang、Y. Wu、X. Xie、Z. Zhou、H. Zhang、X. L. Zhang 和 Y. Wu，《Eventanchor：减少球拍类运动视频事件标注中的人工交互》，载于 CHI 会议论文集。纽约：ACM，2021年。[45] C. Shi、Y. Wu、S. Liu、H. Zhou 和 H. Qu，《Loyaltracker：搜索引擎中的忠诚度动态可视化》，IEEE 可视化与计算机图形学汇刊，第 20 卷，第 12 期，第 1733–1742 页，2014年。[46] B. C. Cappers 和 J. J. van Wijk，《利用规则、聚合和选择探索多变量事件序列》，IEEE 可视化与计算机图形学汇刊，第 24 卷，541，2018 年 1，532，第 – 页。[47] J. Magallanes、T. Stone、P. D. Morris、S. Mason、S. Wood 和 M.-C. Villa- Uriol，《Sequen-c：时间事件序列的多层概览》，IEEE 可视化与计算机图形学汇刊，第 28 卷，第 1 期，第 901–911 页，2022年。[48] Z. Liu、B. Kerr、M. Dontcheva、J. Grover、M. Hoffman 和 A. Wilson，《Coreflow：从事件序列中提取并可视化分支模式》，Computer Graphics Forum，第 36 卷，538，2017 年 3，527，第 – 页。[49] P. J. Polack Jr.、S.-T. Chen、M. Kahng、K. D. Barbaro、R. Basole、M. Sharmin 和 D. H. Chau，《Chronodes：事件序列的交互式多焦点探索》，ACM 交互式智能系统汇刊，第 8 卷，第 1 期，2018年2月。

[50] Y. Guo、S. Guo、Z. Jin、S. Kaul、D. Gotz 和 N. Cao，《事件序列数据可视分析综述》，*IEEE* 可视化与计算机图形学汇刊，第 1–1 页，2021年。[51] P. H. Nguyen、C. Turkay、G. Andrienko、N. Andrienko、O. Thonnard 和 J. Zouaoui，《通过动作序列理解用户行为：从常见到罕见》，*IEEE* 可视化与计算机图形学汇刊，第 25 卷，第 9 期，第 2838–2852 页，2019年。[52] A. Yeshchenko、C. Di Ciccio、J. Mendling 和 A. Polyvyanyy，《面向业务流程序列数据分析的可视化漂移检测》，*IEEE* 可视化与计算机图形学汇刊，第 1–1 页，2021年。[53] S. Guo、Z. Jin、Q. Chen、D. Gotz、H. Zha 和 N. Cao，《通过序列匹配和可视化比较实现事件序列中的可解释异常检测》，*IEEE* 可视化与计算机图形学汇刊，第 1–1 页，2021年。[54] D. Chen、H. Chen、Y. Yang、A. Lin 和 Z. Yu，《基于动作的对话数据集：用于构建更深入的面向任务的对话系统的语料库》，载于 *NAACL-HLT* 会议论文集。线上：ACL，2021年6月，第 3002–3017 页。[55] M. Sedlmair、M. Meyer 和 T. Munzner，《设计研究方法：来自一线实践与文献积累的反思》，*IEEE* 可视化与计算机图形学汇刊，第 18 卷，2440，2012 年 12，2431，第 – 页。[56] D. Qi、W. Tan、Q. Yao 和 J. Liu，《Yolo5face：为何要重新发明人脸检测器》，*ArXiv* 预印本 *ArXiv:2105.12931*，2021年。[57] S. Li 和 W. Deng，《深度面部表情识别：综述》，*IEEE* 情感计算汇刊，第 1–1 页，2020年。[58] T.-Y. Yang、Y.-T. Chen、Y.-Y. Lin 和 Y.-Y. Chuang，《Fsa-net：从单幅图像中学习用于头部姿态估计的细粒度结构聚合》，载于*IEEE/CVF* 计算机视觉与模式识别会议会议论文集，2019年，第 1087–1096 页。[59] K. Wang、X. Peng、J. Yang、S. Lu 和 Y. Qiao，《抑制大规模面部表情识别中的不确定性》，载于 *IEEE/CVF* 计算机视觉与模式识别会议会议论文集，2020年6月。[60] H. Bredin、R. Yin、J. M. Coria、G. Gelly、P. Korshunov、M. Lavechin、D. Fustes、H. Titeux、W. Bouaziz 和 M.-P. Gill，《Pyannote.audio：说话人分离的神经网络构建模块》，载于*IEEE* 声学、语音与信号处理国际会议，2020年，第 7124–7128 页。[61] M. G. de Pinto、M. Polignano、P. Lops 和 G. Semeraro，《结合深度神经网络与梅尔频率倒谱系数，从口语中理解情绪的模型》，载于*2020*年 *IEEE* 进化与自适应智能系统会议 *(EAIS)*，2020年，第 1–5 页。[62] W. Xu、L. Huang、A. Fox、D. Patterson 和 M. I. Jordan，《通过挖掘控制台日志检测大规模系统问题》，载于 *SIGOPS* 会议论文集。美国纽约：ACM，2009年，第 117–132 页。[63] J.-G. Lou、Q. Fu、S. Yang、Y. Xu 和 J. Li，《从控制台日志中挖掘不变量以检测系统问题》，载于*USENIX* 会议论文集，USENIXATC'10 系列。美国：USENIX 协会，2010年，第 24 页。[64] N. Ye，《用于异常检测的时序行为马尔可夫链模型》，载于*IEEE* 系统、人类与控制论信息保障与安全研讨会论文集，第 166 卷，2000年，第 169 页。[65] L. Yang、C. Xiong、J. K. Wong、A. Wu 和 H. Qu，《基于众包入门描述信息可视化的示例讲解与经验总结》，*IEEE* 可视化与计算机图形学汇刊，第 1–1 页，2021年。[66] M. Gleicher，《可视化比较的考量》，*IEEE* 可视化与计算机图形学汇刊，第 24 卷，423，2018 年 1，413，第 – 页。[67] B. Morrow、T. Manz、A. E. Chung、N. Gehlenborg 和 D. Gotz，《用于为异构时间图表提供背景信息的外围图》，载于*2019*年 *IEEE* 可视化会议 *(VIS)*，2019年，第 1–5 页。[68] J. Zhou、X. Wang、J. K. Wong、H. Wang、Z. Wang、X. Yang、X. Yan、H. Feng、H. Qu、H. Ying 和 W. Chen，《Dpviscreator：通过差分隐私将模式约束融入隐私保护可视化》，*IEEE* 可视化与计算机图形学汇刊，第 29 卷，第 1 期，第 809–819 页，2023年。



Kam KwaiWong 是香港科技大学（HKUST）计算机科学与工程系的博士候选人。他在香港科技大学获得工程学学士学位。主要研究兴趣包括数据可视化、可视分析和数据挖掘。详情请参阅 https://jasonwong.vision。



XingboWang 是香港科技大学（HKUST）计算机科学与工程系的博士候选人。他于2018年在中国武汉大学获得工程学学士学位。研究兴趣包括人机交互（HCI）、数据可视化、自然语言处理（NLP）和多模态分析。详情请参阅 https://andy-xingbowang.com。



Yong Wang 是新加坡管理大学计算与信息系统学院的助理教授。他的研究兴趣包括数据可视化、可视分析和可解释机器学习。他于2018年获得香港科技大学计算机科学博士学位。他分别获得哈尔滨工业大学和华中科技大学的工程学学士和硕士学位。更多详情请参阅 http://yong-wang.org。



Jianben He 是香港科技大学（HKUST）计算机科学与工程系的博士生。她获得了华中科技大学电子信息与通信专业的工程学学士学位。更多详情请参阅 https://jessiehe970311.github.io/。



Rong Zhang 是香港科技大学跨学科项目办公室的讲师。他的研究兴趣包括数据可视化、以人为本的设计和跨学科教育。



HuaminQu 是香港科技大学计算机科学与工程系（CSE）讲席教授、跨学科项目办公室（IPO）主任及新兴跨学科领域分部（EMIA）主任。他在中国西安交通大学获得数学理学学士学位，并在石溪大学获得计算机科学理学硕士和博士学位。他是 VisLab 主任，同时还担任人机交互（HCI）组在计算机科学与工程系的协调员。他的主要研究兴趣是可视化和人机交互，重点关注城市信息学、社会网络分析、电子学习、文本可视化和可解释人工智能（XAI）。