doi:10.3969/j.issn.1001-4616.2026.05.013

基于终端位姿感知的多模态视频抽稀算法

杨静丽[1],戴佳维[2],唐钱进[2],马燕芹[1]

(1.南京工业职业技术大学,江苏 南京 210023)
(2.南京魔数团信息科技有限公司,江苏 南京 210012)

[**摘要**] 针对第一视角长时视频冗余信息多、传统抽帧难以捕获语义关键帧的问题,本文提出一种基于终端设备位姿感知的多模态大模型视频抽稀与分析算法. 该方法充分利用同步获取的惯性测量单元数据、音频及视觉内容,实现多触发协同的关键帧抽取机制,通过位姿突变、语音语义触发和视觉显著性综合判定,精准捕获关键事件. 在此基础上,将关键帧组织为语义片段,与位姿统计、语音转写等元信息绑定,形成适配大模型输入的"图像序列+文本描述"结构. 端侧执行轻量抽稀,云侧大模型进行推理分析,实现高效协同. 在自建工业端子排巡检数据集上的实验结果表明,该方法在关键帧检测准确率、语义覆盖度、压缩比及实时性上均优于对比方法. 该研究为长时第一视角视频智能分析提供了高效、可部署的工程化方案.

[**关键词**] 视频抽稀,多模态大模型,终端位姿感知,第一视角视频
[**中图分类号**]TP391　[**文献标志码**]A　[**文章编号**]1001-4616(2026)05-0119-11

Multimodal Video Sparsification Algorithm Based on Terminal Pose Perception

Yang Jingli[1], Dai Jiawei[2], Tang Qianjin[2], Ma Yanqin[1]

(1.Nanjing University of Industry Technology, Nanjing 210023, China)
(2.Nanjing Magic Cube Information Technology Company Limited, Nanjing 210012, China)

Abstract: To address the issue that long first-person videos contain a lot of redundant information and traditional frame sampling struggles to capture key semantic moments, this paper proposes a multimodal large-model-based video sparsification and analysis method driven by terminal device pose awareness. The method makes full use of synchronously collected data, including inertial measurement unit data, audio and visual content. It uses a multi-trigger collaborative keyframe extraction mechanism, combining sudden pose changes, speech-based semantic triggers and visual saliency to accurately capture important events. On top of that, the extracted keyframes are organized into semantic segments, which are linked with metadata such as pose statistics and speech transcriptions. These are then structured into a "image sequence + text description" format suitable for large model input. The system performs lightweight video thinning on the device side, while the cloud-based large model handles inference and analysis, enabling efficient collaboration. Experimental results on a self-built industrial terminal block inspection dataset show that this method outperforms baseline approaches in keyframe detection accuracy, semantic coverage, compression ratio, and real-time performance. This study provides an efficient and deployable engineering solution for intelligent analysis of long-duration first-person videos.

Key words: video sparsification, multimodal large mode, terminal pose perception, first-person video

随着智能巡检、执法记录、工业运维及增强现实等应用场景的快速发展,基于可穿戴终端或移动设备采集的第一视角视频数据量持续增长[1]. 与传统固定摄像头视频相比,第一视角视频直接反映用户做任务时的观察对象、操作行为和环境变化,因此对行为理解、事件回溯、任务辅助、状态监测等具有重要应用价值. 尤其在复杂动态环境下,第一视角视频能够更真实地记录作业流程与交互细节,为后续智能感知与

收稿日期:2026-03-13.
基金项目:国家自然科学基金资助项目(62503220)、江苏省自然科学基金资助项目(BK20230706).
通讯作者:马燕芹,博士,副教授,研究方向:具身智能、智能机器人、计算机视觉技术. E-mail:mayanqin@niit.edu.cn

决策提供丰富的信息基础. 但是这类视频跟常规视频数据有明显不同:首先,采集时间长,原始数据量大,存储、检索和传输成本高;其次,视角随身体运动不断变化,视频里常出现场景切换、抖动、光照变化和遮挡干扰,时空连续性差[2]. 在长时间采集中真正跟任务相关的关键事件、重要目标或有意义的信息往往只在少量片段里出现,大量内容重复、冗余或没价值,语义稀疏性很明显. 这些特点给视频压缩存储、实时传输、关键片段筛选和智能分析都带来不小挑战[2].

近年来,可穿戴设备和第一视角视频的多模态智能分析逐渐受到关注. Hong 等[3]提出 EgoLM 模型,把第一视角视频与运动传感器信息联合建模,用于运动理解和语言描述生成. 这说明在可穿戴场景下,运动状态和传感器信号能为行为理解提供有用的补充语义. 也就是说,对于第一视角长视频,仅依靠视觉内容往往不容易说清用户真正的关注点和动作意图,加入位姿或运动相关信息已成为提升场景理解能力的重要方向.

因此,如何从第一视角长视频中抓取关键、去除冗余并提升处理效率,是一个重要问题. 现有方法多按固定帧率或画面变化提取关键帧,虽能减少冗余[4],但未利用终端多模态信息,难以准确刻画操作者意图. 多模态大模型对输入规模与语义密度高度敏感,传统抽帧在成本与效率上的局限因此更为突出[5]. 这类模型虽能提升语义理解,但高度依赖输入质量[6]. 为此,在保证语义完整性的前提下实现高效抽稀,成为视频理解的关键[7]. 对于长视频太冗长,关键信息容易丢失的问题,Tang 等[8]采用自适应关键帧采样方法,在固定视觉 token 预算下既考虑关键帧与任务提示的相关性,也顾及视频内容覆盖度,以此提升理解效果. Lee 等[9]则基于大语言模型做视频摘要:先用多模态模型生成帧级描述,再用语言模型在局部和全局上下文里评估帧的重要性,最后得到语义更一致的摘要. 可见,长视频理解中的关键帧选择与语义压缩正成为多模态大模型应用的关键环节,也说明在前端构建高语义密度、低冗余的视频表示很有必要.

现有视频抽稀方法主要分三类:基于规则的方法、基于深度学习的方法、基于多模态建模的方法. 传统方法大多靠帧间差异或信息量(如熵、互信息),计算开销虽低,但很难抓住高层语义[10-11];深度学习方法通过注意力或时序建模提升语义表达能力,但计算复杂度通常较高[12];近年来,多模态大模型方法逐渐兴起,融合视觉与语言信息显著提升视频理解能力,但对输入数据规模与语义密度要求较高[13-14]. 多模态大模型在视频理解中展现出较强的语义建模能力,如相关模型已在跨模态对齐与推理任务中取得进展[15]. 但这类模型对输入规模与语义密度较为敏感,直接处理长时视频成本较高,通常依赖关键帧或结构化片段作为输入[16-17]. 因此,在计算约束下构建“高语义密度、低冗余”的视频表示成为关键问题. 基于此,有必要设计一种面向终端部署的多模态视频抽稀方法,在降低数据规模的同时提升语义表达效率,为后续大模型分析提供更优输入.

针对上述问题,本文提出一种基于终端位姿感知的多模态视频抽稀方法,用采集过程中同步获得的位姿、音频和视觉信息,把位姿变化、语音触发和视觉显著性统一成多触发协同判据,自适应选取关键帧;终端侧完成初步抽稀,为云端提供语义更集中的输入. 在此基础上,本文设计了一种面向多模态大模型的视频语义结构化表示方法,按时间连续性和语义一致性将关键帧组织成结构化片段,再与对应的位姿状态、时间戳和语音文本关联建模. 这种表示保留事件时序和空间变化,能适配多模态大模型的输入需求,提升模型对复杂动态场景的理解. 通过解耦终端侧多模态感知与云端大模型推理[18],在保证分析准确性的同时控制计算开销与延迟,适用于长时连续第一视角视频的智能分析.

本方法主要面向移动终端场景,例如工业巡检、执法记录及增强现实辅助操作等. 在此类场景中,视频通常由可穿戴设备或手持设备采集,具有第一视角、长时连续及语义稀疏等特点. 本方法作为前端视频处理模块,主要承担语义抽稀与结构化表示任务,为后续大模型分析提供高质量输入. 本文主要创新点:

(1)融合终端位姿、语音和视觉信息的多模态视频抽稀方法,通过建模人类关注行为,提升关键帧选取的语义相关性;

(2)设计多触发协同的抽稀机制,自适应捕获长视频中的关键事件;

(3)构建面向多模态大模型的结构化视频语义表示,降低输入冗余,提升视频理解与分析效率.

1 基于终端位姿感知的多模态视频抽稀方法

本节提出一种“终端侧轻量抽稀-云端大模型分析”的工程化方法. 该方法在终端设备上结合位姿、音

频与视觉信息实现多触发关键帧抽取,并将结果整合为结构化语义片段,最后通过统一接口提供给多模态大模型进行高层语义分析.

1.1 方法整体框架

本文方法面向可穿戴摄像设备、智能手机和执法记录仪等移动终端,设备具备摄像头、惯性测量单元IMU(Inertial Measurement Unit)和麦克风. 终端侧进行低成本抽稀,云端大模型完成语义推理,实现实时性与分析能力的平衡. 终端采集的数据可表示为按时间戳对齐的多模态序列. 为保证多模态数据对齐精度,这里采用"统一时间戳+软件对齐"的同步策略. 硬件上各传感器数据统一绑定系统时间戳;软件上通过插值与滑动窗口对齐方法实现不同采样率数据的时间同步,例如将 IMU 数据通过线性插值映射到视频帧时间轴,同时对音频数据进行短时分帧并对齐到对应时间区间. 此外,通过估计各模态间的固定延迟偏移(Δt)进行校正,从而实现近似毫秒级同步精度.

抽稀决策模块由多触发判据和静默补偿策略构成,主要包括:用于捕捉操作者注意力转移,例如转头、俯仰或快速移动导致的视角变化的位姿突变触发;捕捉与任务相关的口述信息,如报读编号、异常描述或指令交互的语音语义触发;识别画面中出现的新信息或结构性变化的视觉显著性触发;在无触发情况下保持低频采样,以保证事件完整性及时间连续性的静默补偿. 终端侧输出不仅仅是离散关键帧集合,更是由若干语义片段组成的结构化表示. 每个片段都包括关键帧序列、时间窗信息、片段内位姿统计特征、语音转写及触发原因标签,以便于后续大模型进行分段推理与跨段汇总. 云端侧以结构化片段为输入,采用逐片段推理和全局汇总的两级处理流程:首先在每个片段内部完成对象或事件描述与局部推理,然后在全局层面对时间线进行归纳、对比与结论生成,从而降低长视频直接输入大模型带来的上下文压力与成本.

图 1 展示了方法整体框架的分层结构,包括数据采集层、抽稀决策层、语义分析层. 数据采集层包含 IMU 位姿数据、原始视频帧与音频流;抽稀决策层由位姿触发、语音触发、视觉触发及静默补偿等子模块

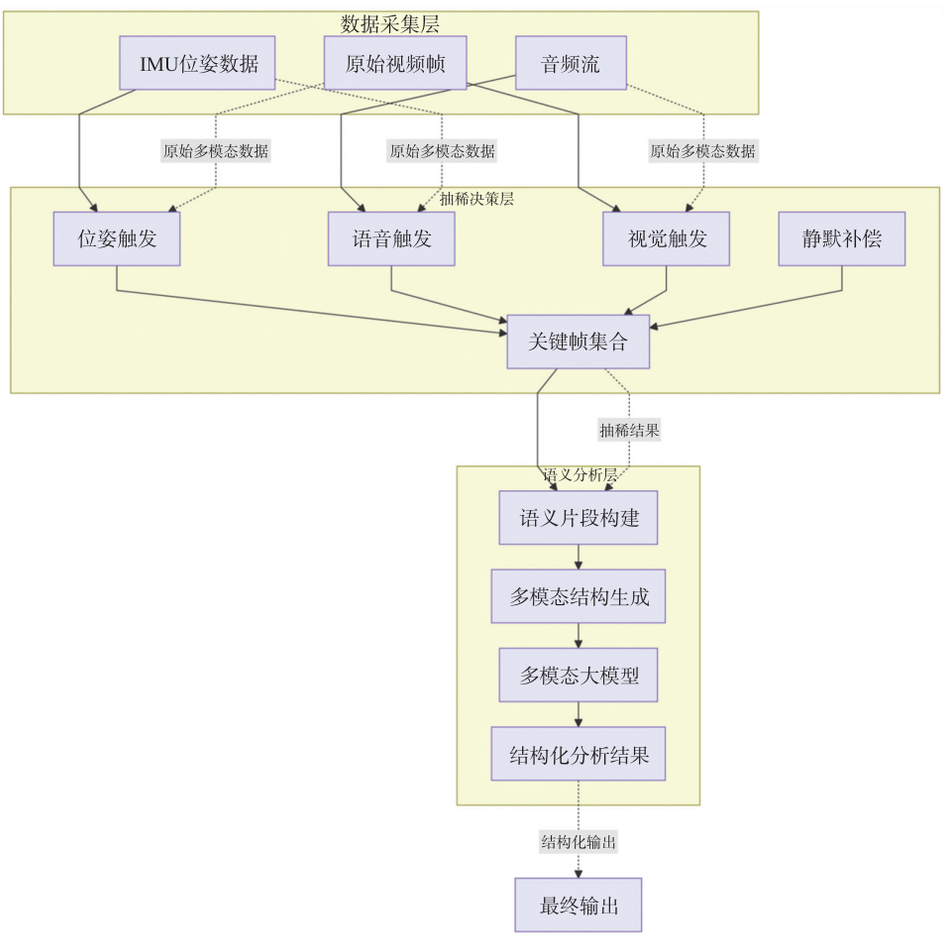


图 1 基于终端位姿感知的多模态视频抽稀整体框架

Fig. 1 Overall framework of multimodal video thinning based on terminal pose awareness

构成,输出关键帧集合;语义分析层将关键帧组织为语义片段,生成“图像序列+文本元数据”结构输入到多模态大模型. 图中箭头标注了不同类型数据流,包括原始多模态数据、抽稀结果、结构化分析输出.

本算法中,语义片段在端云之间的稳定传输是算法落地部署的重要前提. 但在实际部署中,终端设备是处于移动网络环境,所以本文设计了面向不稳定网络的鲁棒传输机制. 终端侧使用基于时间窗口的环形缓冲区对抽稀后的语义片段进行缓存,并按照“关键语义优先”原则进行调度传输,其中包含语音触发片段或异常事件片段的数据优先级最高. 在数据传输过程中,为应对网络中断情况,采用分片上传与断点续传机制. 此外,在网络不可用时,终端侧启用轻量本地分析模块,对关键片段进行初步处理,并在网络恢复后进行延迟同步,从而保证系统整体运行的连续性与可靠性. 上述机制主要从工程部署角度保证结构化语义片段能够稳定送达云端. 在此基础上,对终端侧参与抽稀决策的多模态感知信息形式化建模.

1.2 多模态感知信息建模

将视频表示为 $V=\{I_t\}_{t=1}^{T}$,其中 $I_t\in\mathbf{R}^{H\times W\times 3}$ 表示时间戳 t 对应的图像帧. 为适应终端侧轻量计算,本文未引入复杂的 3D CNN(Convolutional Neural Network),视觉特征仅需反映画面变化强度或新信息出现的低维指标,用于抽稀触发判定与静默补偿采样.

终端位姿状态表示为 $P=\{p_t\}_{t=1}^{T}$,其中 $p_t=(q_t,\omega_t,a_t)$,由姿态四元数 q_t、角速度 ω_t 和线加速度 a_t 组成. 工程实现中,可使用传感器融合算法输出的姿态角即滚转、俯仰、偏航及角速度模长等指标. 本文重点关注位姿变化率及短时突变,而非绝对位姿值,以捕捉操作者注意力的快速转移.

音频流表示为 $A=\{x_t\}_{t=1}^{T}$,其中 x_t 为对应时间戳的短时音频片段. 在音频处理中,首先进行语音活动检测,定位有效语音区间,再通过自动语音识别,获得包含起止时间信息的转写文本 S,S 是由一串按顺序的输出单元 s_k 组成的序列,即 $S=[s_1,s_2,\cdots s_k\cdots]$,其中 s_k 是转写出来的第 k 个词,S 是最终转写结果. 在抽稀阶段,本文主要利用两类信号,一是用于触发“讲话即关键”的采样策略的语音存在性;二是通过关键词或任务词典、轻量文本分类器识别高价值语句,如异常、告警、编号或操作指令的语义强度.

综上,终端侧多模态同步输入可形式化为:

$$X=\{(I_t,p_t,x_t)\mid t=1,2,\cdots,T\}. \tag{1}$$

使用统一表示,多触发抽稀策略就可在同一时间轴上完成判定与信息融合,同时避免仅依赖视觉特征导致的语义遗漏.

1.3 基于多触发机制的关键帧抽稀策略

1.3.1 设计动机与总体判定

第一视角长视频的冗余可以归结为两类:一是视角长时间停留而导致的大量相似帧;二是终端设备的自然抖动引起的“伪变化”. 为此,本文提出一种组合抽稀策略,即“强语义触发优先、弱视觉变化辅助、位姿突变纠偏、静默补偿兜底”. 对每个时间戳 t,定义综合触发得分:

$$\sigma_t=w_p\sigma_t^{(p)}+w_a\sigma_t^{(a)}+w_v\sigma_t^{(v)}, \tag{2}$$

其中,$\sigma_t^{(p)}$、$\sigma_t^{(a)}$、$\sigma_t^{(v)}$ 分别对应位姿、语音与视觉触发分量,w_p、w_a、w_v 为对应权重参数. 在权重设定方面,本文采用“经验初始化+数据驱动优化”的策略:初始权重依据任务先验进行设定,通常语音触发权重较高,其次为位姿与视觉;随后在验证集上以语义覆盖度为目标,通过贝叶斯优化进行调优. 在扩展实现中,权重通过轻量神经网络进行动态预测,实现对不同场景的自适应调节. 当 $\sigma_t\geqslant\tau$ 时,选取该帧为关键帧;否则进入静默补偿逻辑.

1.3.2 位姿突变触发机制

在交互过程中,当操作者关注对象发生转移时,终端设备通常会出现角速度峰值或姿态角快速变化,把该信息作为语义切换点的强指示信号. 为此,在这里定义短时窗口 $[t-t_\Delta,t]$,内角速度模长均值为 $\overline{\|\omega\|}$,当前时刻角速度模长为 ω_t,则位姿触发分量定义为:

$$\sigma_t^{(p)}=\max\left(0,\frac{\omega_t-\overline{\|\omega\|}}{\overline{\|\omega\|}+\varepsilon}\right). \tag{3}$$

快速转头或俯仰等行为下,$\sigma_t^{(p)}$ 增大并触发采样. 该机制能抑制纹理轻微变化但语义不变时的冗余采样,也减少仅依赖帧差带来的误触发. 例如,巡检时,操作者从一个设备快速转向另一个设备,或低头看标签信息时,IMU 产生明显角速度峰值,系统随即触发关键帧采集,捕捉语义切换点.

1.3.3 语音语义触发机制

巡检、执法及运维等典型人机协作场景中,操作者常通过语音描述来记录关键对象. 相比视觉信息,语音通常更早、更明确地反映任务的语义内涵. 基于此,对每个语音片段 s_k 及其时间窗 $[t_k^s, t_k^e]$,设计一个语音触发机制,只要满足任一条件,就给该区间内的视频帧打上语音触发标记.

(1)语音活动检测判定为有效语音且持续时长超过阈值,排除短时噪声;

(2)语音转写文本中包含了任务相关关键词,例如“异常”、“破损”、“泄漏”、“告警”、“编号”、“完成”、“开始”等;

(3)文本分类器判定语音转写文本为高价值语句,如异常描述、操作确认、现场口述记录.

基于上述条件,语音触发分量 $\sigma_t^{(a)}$ 可定义为二值标记或连续语义置信度值. 加入这一机制后,关键帧抽稀的语义相关性明显提高,特别适合视觉变化小但语义信息多的场景.

1.3.4 视觉显著性触发机制

机器人或智能设备在环境交互中,只靠本体感知比如位姿变化、语音指令,这类显式触发机制,很难捕捉到环境里非交互性的关键动态. 当位姿和语音触发不显著时,仍需要有效检测画面中新对象出现、光照突变和场景切换等视觉事件. 为此,本文提出一个基于轻量视觉变化指标的触发机制,作为对前述触发的补充.

终端设备采用计算开销较小的视觉差异度量方法来衡量较连续图像帧的变化程度. 度量指标包括相邻帧灰度直方图距离(定义为 $D_{\text{hist}} = \sum_k |H_t(k) - H_{t-1}(k)|$)、下采样图像的像素差分能量 $\left(\text{定义为 } E_t = \frac{1}{N} \times \sum_{i=1}^{N} (I_t(i) - I_{t-1}(i))^2\right)$、全局特征向量的余弦距离等. 度量结果记作变化度 d_t,用于后续关键帧选取,并有如下定义:

$$\sigma_t^{(v)} = \min\left(1, \frac{d_t}{\tau_v}\right), \tag{4}$$

其中,τ_v 为视觉变化阈值.

1.3.5 静默补偿策略

长时间无事件触发时,完全停止数据采样会打断环境感知时间序列,破坏视觉时间连续性,削弱大语言模型对事件演进过程的理解. 在连续未触发事件的时间段内,系统按最小采样间隔 K 补偿抽帧. K 取值与场景动态性相关:运动频繁或语义变化快的场景取较小 K,保证时间分辨率;静态场景取较大 K,提高压缩效率. 工程实现中,依据短时位姿变化强度或视觉变化幅度对 K 进行自适应调整. 补偿帧作为后续语义片段中的时间锚点,维持事件连续性,辅助大模型全局推理.

1.4 结构化视频语义片段构建

1.4.1 片段定义与边界确定

基于触发关键帧,为提升信息表征的结构化水平,依据时间邻近性与触发语义对关键帧进行聚合,形成具有完整语义边界的视频片段,满足大语言模型对上下文理解与输入负载的要求. 设最终保留的关键帧集合为:

$$\mathcal{F} = \{(t_i, I_{t_i})\}_{i=1}^{N}, \tag{5}$$

其中,t_i 为关键帧时间戳. 基于触发类型及时间邻近性,将关键帧聚合为语义片段的集合:

$$\mathcal{L} = \{g_j\}_{j=1}^{M}. \tag{6}$$

每个片段 g_j 对应连续时间窗 $[T_j^s, T_j^e]$,其边界由如下规则确定:

(1)时间间隔约束:相邻关键帧时间间隔超过阈值 ΔT,则划分为不同片段;

(2)强语义对齐约束:位姿突变或高价值语音触发时,可强制开启或终止片段,使语义边界与行为转折点一致;

(3)长度上限约束:设定片段最大时长或最大关键帧数量,防止单片段过长使后续大模型输入负载过高.

1.4.2　片段级多模态绑定

为提升片段表示的可解释性与可检索性,对每个片段 g_j 绑定多模态元信息,包括:时间信息片段,即起止时间 T_j^s, T_j^e 及关键帧时间戳序列;空间与运动统计,包含平均角速度、最大姿态变化幅度及运动平稳度等位姿变化统计量;语音语义信息,包含 ASR(Automatic Speech Recognition)转写文本及关键词匹配结果;触发机制标签,包含 pose、voice、vision、compensation 等触发类型及对应置信度;视觉摘要信息,主要选取首帧、中位帧或触发峰值帧作为代表帧,以支持快速语义定位与检索.

1.5　结构化输出流程

在完成语义片段构建及多模态元信息绑定后,还需将片段组织为适配多模态大模型的统一输入形式,并定义相应的分析流程. 设语义片段集合为 $G=\{g_j\}_{j=1}^{M}$,任务指令集 $T=\{\tau_{\text{task}}, \tau_{\text{constraint}}, \tau_{\text{format}}\}$,其中 τ_{task} 表示具体业务目标,例如异常识别、事件复核、流程核对等,$\tau_{\text{constraint}}$ 表示输出约束条件,τ_{format} 表示结果格式规范. 视频分析接口函数定义为:

$$\mathcal{R}=\Phi(G,\mathcal{J}), \tag{7}$$

其中,Φ 为多模态大模型推理算子,$\mathcal{R}$ 为结构化分析结果.

在实现层面,Φ 采用分层结构:片段级推理函数 ϕ_{local} 与全局汇总函数 ϕ_{global}:

$$\mathcal{R}=\phi_{\text{global}}(\{\phi_{\text{local}}(g_j,\mathcal{T})\}_{j=1}^{M}), \tag{8}$$

其中,$r_j=\phi_{\text{local}}(g_j,\mathcal{T})$ 输出关键对象、事件描述、证据桢索引及置信度;$\phi_{\text{global}}(\{r_j\}_{j=1}^{M})$ 则完成实体对齐、事件合并、时间线组织与报告生成.

为保证输入稳定性,定义编码函数 $z_j=\Psi(g_j)$,将片段统一表示为“图像序列+文本元数据”格式. 最终输出空间定义为:

$$\mathcal{R}=\{(E_k,\mathcal{E}_k)\}_{k=1}^{K}. \tag{9}$$

并建立证据映射关系 $\mathcal{E}_k\subseteq\{(j,t_{j,i},\ell_{j,i})\}$,确保分析结果可追溯至原始片段与关键帧.

算法实现过程的伪代码如算法 1 所示. 算法以多模态同步流 $X=\{(I_t,p_t,x_t)\}_{t=1}^{T}$ 作为输入,通过逐帧触发判定完成关键帧筛选,并在此基础上构建语义片段进行分层推理,最终输出结构化结果 $\mathcal{R}$.

算法 1　基于终端位姿感知的多模态视频抽稀与分析

Require:多模态同步视频输入 $X=\{(I_t,p_t,x_t)\}_{t=1}^{T}$,任务指令集合 T
Ensure:结构化分析结果 $\mathcal{R}$
初始化关键帧集合 $\mathcal{F}\leftarrow\phi$
for 每个时间戳 $t=1$ 到 T do
　计算位姿触发分量 $\sigma_t^{(p)}$
　计算语音触发分量 $\sigma_t^{(a)}$
　计算视觉触发分量 $\sigma_t^{(v)}$
　计算综合触发得分 $\sigma_t=w_p\sigma_t^{(p)}+w_a\sigma_t^{(a)}+w_v\sigma_t^{(v)}$
　if $\sigma_t\geqslant\tau$ then
　　将 I_t 加入关键帧集合 $\mathcal{F}$
　　记录触发类型与时间戳
　else
　　按静默补偿策略决定是否保留帧
　end if
end for
将关键帧集合 $\mathcal{F}$ 聚合为语义片段集合 $\mathcal{G}=\{g_j\}_{j=1}^{M}$
for 每个片段 g_j do
　编码片段为统一输入格式 $z_j=\Psi(g_j)$
　局部推理 $r_j=\phi_{\text{local}}(z_j,T)$
end for
全局汇总 $\mathcal{R}=\phi_{\text{global}}(\{r_j\}_{j=1}^{M})$
return $\mathcal{R}$

2 实验数据与结果分析

2.1 数据集

本文使用第一视角可穿戴终端(参数见表 1)录制了 10 段工业端子排巡检视频,总时长约 65 min,内容覆盖正常巡检、异常检测及干扰场景(端子排样式见图 2). 数据集统计信息见表 2. 为构建客观基准,建立三层真值标注体系:视觉层面由两名专家独立标注关键帧并取交集;语义层面标注语音触发点及异常类别;运动层面标注 IMU 姿态突变点及运动峰值区间. 专家标注一致性结果见表 3,表明标注具有较高可靠性,故将该交集作为后续对比实验的真值. 工业巡检中采集视频数据,用于检查端子排状态并识别异常. 佩戴式设备从第一视角录制,人工巡检包括正常操作、异常检测与干扰行为.

表 1 数据采集系统配置参数

Table 1 Configuration parameters of the data acquisition system

参数类别	具体项目	配置/规格
硬件平台	采集设备	Vuzix M400 智能眼镜
视频参数	分辨率	1 920×1 080(全高清)
	帧率	60 fps
	编码格式	H.264
	像素格式	NV12(YUV 4:2:0 采样)
惯性参数	IMU 采样率	100 Hz
	姿态表示	欧拉角(横滚角、俯仰角、偏航角)
音频参数	音频采样率	16 kHz

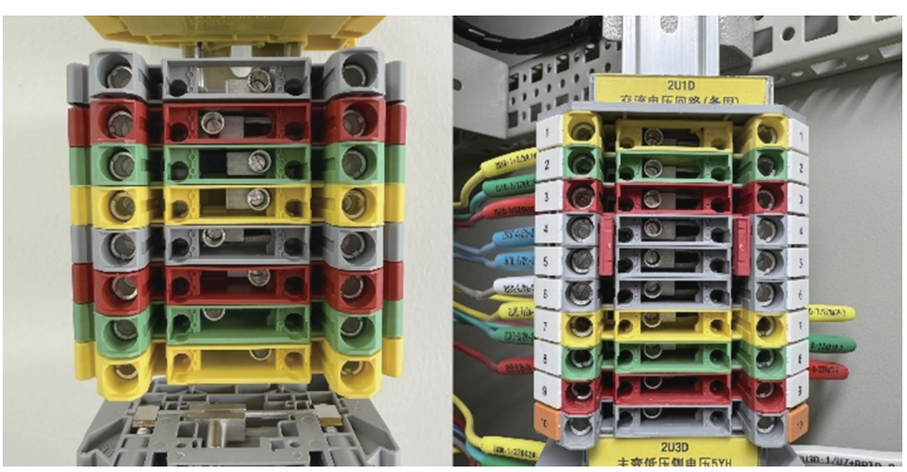


图 2 端子排示例图

Fig. 2 Terminal strip example diagram

表 2 数据集统计信息

Table 2 Dataset statistics

项目	数值
视频数量	10
总时长	约 65 min
平均时长	6.5 min
异常视频数量	6
正常视频数量	2
干扰场景视频数量	2
异常事件总数[1]	87
视频总帧数	234 891
参考关键帧总数[2]	73 414

注:[1] 异常事件包括:端子松动 34 次、状态错误 28 次、污渍痕迹 3 次、标签模糊或破损 22 次;[2] 参考关键帧总数为本文实验中用于实验评估的 Ground Truth 关键帧数量,与专家标注的关键帧无关.

表 3 工业巡检专家标注结果

Table 3 Industrial inspection expert annotation results

	专家 1	专家 2
标注关键帧总数	78 528	72 904
命中参考关键帧数量	68 092	64 892
False Positive 帧数量	10 436	8 012
False Negative 帧数量	5 322	8 522
Precision	0.867	0.890
Recall	0.927	0.884
F1-score	0.896	0.887
Accuracy	0.933	0.931
Cohen's Kappa	0.847	0.839

为说明抽稀结果在下游巡检语义分析中的支撑作用,图 3 给出了电气柜端子区域的检测与语义标注结果. 可以看出,本文方法能够在复杂现场背景下较为准确地定位端子、设备标牌及其对应文本区域,并对关键语义目标进行有效区分. 图中不同颜色框分别表示不同类别的检测目标与语义关注区域,表明所构建的多模态感知与语义关联机制能够实现对巡检场景中关键对象的稳定识别与信息对齐,从而为后续关键帧筛选及语义片段构建提供可靠支撑.



图 3　电气柜端子及标牌检测结果示意图

Fig. 3　Illustration of terminal and nameplate detection results in an electrical cabinet

2.2　对比实验设置及评价指标

为验证本文方法的有效性,选取五种代表性基线:Entropy+MI(基于信息熵与互信息的滑动窗口方法)、Optimal Frame(基于最优子集选择的关键帧提取方法)、Attention-DNN(基于注意力机制的深度神经网络方法)、LMSKE 及 TriPSS(基于多模态语义建模的关键帧提取方法). 不同于上述基线,本文方法采用多模态融合策略,评估其在关键帧选取任务上的性能.

从准确性、压缩效率、语义完整性及实时性四个维度评价算法性能. 其中,准确性使用精确率(Precision)、召回率(Recall)和 F1 分数度量;压缩效率用压缩比(原始总帧数/输出关键帧数)衡量;语义完整性通过冗余帧比例(输出中高度相似帧占比)和语义覆盖度(关键帧对事件内容的覆盖程度)评估;实时性通过平均推理时间和是否达到 60 fps 来衡量. 语义覆盖度定义为:

$$\text{Coverage}=\frac{N_{\text{covered}}}{N_{\text{total}}}, \tag{10}$$

其中,N_{total}为数据集中标注的语义事件总数,N_{covered}为关键帧成功覆盖的事件数. 具体实现时,关键帧时间戳落入事件时间窗口内,或其语义特征与标注事件满足相似度阈值,即认为被覆盖.

2.3　实验结果

实验视频数据的原始平均帧数为每视频 23 489 帧. 算法性能对比实验在 NVIDIA Jetson Orin NX(16 GB 内存版)平台上进行. 表 4—表 7 汇总了各项实验的对比结果.

表 4　关键帧检测准确性对比

Table 4　Keyframe detection accuracy comparison

方法	Precision	Recall	F1-score	语义覆盖度
Entropy+MI	0.647	0.592	0.618	0.571
Optimal Frame	0.673	0.654	0.663	0.636
Attention-DNN	0.846	0.701	0.767	0.742
LMSKE	0.861	0.812	0.836	0.824
TriPSS	0.803	0.868	0.834	0.857
本文方法	0.907	0.881	0.894	0.872

表 4 显示,各方法在关键帧检测任务上的性能差异明显. 传统方法 Entropy+MI 与 Optimal Frame 等的性能整体偏弱,说明仅靠低层特征难以表征第一视角视频中复杂稀疏的关键语义信息. 深度学习方法 Attention-DNN、LMSKE 和 TriPSS 各项指标都有提升,表明深层语义特征建模能力更强. TriPSS 的 Recall 和语义覆盖度较好,但 Precision 偏低,存在一定误检. 本文方法的 Precision、Recall、F1-score 和语义覆盖度分别为 0.907、0.881、0.894 和 0.872,均为最优,说明该方法既提升了关键帧检测准确性,也兼顾了语义信息的完整覆盖,综合性能更优.

表 5　抽稀效率与冗余控制对比

Table 5　Thinning efficiency and redundant control comparison

方法	平均关键帧数	压缩比	冗余帧比例
Entropy+MI	8 124.600	2.890∶1	0.336
Optimal Frame	7 642.300	3.073∶1	0.301
Attention-DNN	5 984.200	3.924∶1	0.224
LMSKE	5 431.800	4.323∶1	0.198
TriPSS	6 150.400	3.818∶1	0.176
本文方法	4 872.900	4.821∶1	0.139

表 5 显示,各方法在抽稀效率与冗余控制方面差异明显. Entropy+MI 和 Optimal Frame 保留的平均关键帧数偏多,压缩比较低,对冗余信息抑制能力有限. 深度学习方法整体更好,LMSKE 和 TriPSS 已具备较好抽稀能力. 本文方法平均关键帧数最低(4 872.900),压缩比为 1∶4.821,冗余帧比例为 0.139,均优于对比方法,说明该方法在保留信息的同时能更有效压缩冗余,综合性能更好.

表 6 给出了各方法的单帧推理时间与实时性评价. 可以看出,本文方法的推理时间为 19.684 ms/frame,满足实时处理要求. 相比之下,Entropy+MI 与 Optimal Frame 虽然推理速度更快,但二者为基于统计量的轻量启发式方法,结合表 4 的精度结果可知其抽稀质量明显低于本文方法;Attention-DNN 的推理时间为 31.274 ms/frame,已接近实时性临界值;而 LMSKE 与 TriPSS 的推理时间分别为 126.518 ms/frame 与 142.337 ms/frame,无法满足实时需求. 本文方法兼顾检测性能与计算效率,工程应用价值较好.

表 6　计算效率与实时性能对比

Table 6　Computational efficiency and real-time performance comparison

方法	推理时间/(ms/frame)	实时性评价
Entropy+MI	4.963	实时
Optimal Frame	7.842	实时
Attention-DNN	31.274	临界
LMSKE	126.518	慢
TriPSS	142.337	慢
本文方法	19.684	实时

表 7　多模态触发能力对比

Table 7　Multi-modal triggering capability comparison

方法	语音触发捕获率	姿态触发捕获率	连续异常覆盖率
Entropy+MI	0.294	0.238	0.417
Optimal Frame	0.563	0.482	0.605
Attention-DNN	0.702	0.548	0.729
LMSKE	0.781	0.611	0.802
TriPSS	0.863	0.664	0.846
本文方法	0.918	0.889	0.903

表 7 显示,各方法在语音触发捕获率、姿态触发捕获率和连续异常覆盖率上存在明显差异. Entropy+MI 三项指标均较低,Optimal Frame 分别达到 0.563、0.482 和 0.605,优于 Entropy+MI,但对关键触发事件和连续异常片段的覆盖仍较有限. 深度学习方法整体表现更好,其中 TriPSS 三项指标分别为 0.863、0.664 和 0.846. 本文方法分别达到 0.918、0.889 和 0.903,均高于对比方法,相较于 TriPSS,分别提高 5.5、22.5 和 5.7 个百分点,其中姿态触发捕获率提升最为明显. 上述结果表明,本文方法对语音触发事件、姿态突变事件及连续异常片段具有更好的覆盖能力,支持了多模态协同抽稀策略的有效性.

为验证各模态对性能的贡献,设计如下消融实验,结果如表 8 所示. 随着输入模态的逐步增加,Precision、Recall、F1-score 及语义覆盖度稳步提升,说明不同模态信息能够形成有效互补. 仅用视觉模态时,模型具备基础场景表征能力,但性能有限;引入 IMU 后,各项指标明显改善,运动状态信息有助于增强时序动态建模与关键行为感知. 融合语音模态增益更显著,语音信息提供更直接的高层语义线索. 融合视觉、IMU 与语音三种模态后,模型取得最优结果,验证了多模态融合策略在提升识别准确性和语义完整性方面的有效性.

表 8　消融实验结果

Table 8　Ablation experiment results

模型版本	Precision	Recall	F1-score	语义覆盖度
仅视觉	0.765	0.798	0.781	0.736
视觉+IMU	0.823	0.862	0.842	0.804
视觉+语音	0.846	0.876	0.861	0.829
视觉+IMU+语音	0.907	0.881	0.894	0.872

2.4　实验结果分析

通过在工业端子排巡检数据集上的系统对比实验,可以得出以下结论:传统规则方法在视觉抖动场景下易引入冗余关键帧;深度学习方法在语义识别上性能优于传统规则方法,但计算成本较高;现有多模态大模型虽在语义表征能力上具有优势,但直接处理长视频计算开销大,难以满足终端实时部署要求;相比之下,本文提出的多模态融合方法在语义覆盖度、冗余控制与实时性能之间取得了更优的平衡. 此外,消融实验结果表明,IMU 与语音模态对提升关键帧语义捕获能力具有显著作用.

3　结论

针对长时第一视角视频中冗余信息多、传统抽帧方法难以兼顾语义完整性与实时性的难题,本文提出了一种基于终端位姿感知的多模态视频抽稀与分析方法. 该方法融合 IMU 位姿、语音与视觉信息,构建了多触发协同的关键帧抽取机制,并进一步形成面向多模态大模型的结构化语义片段表示,从而实现“终端侧轻量抽稀、云端侧高层分析”的协同处理. 在自建工业端子排巡检数据集上的实验结果表明,本文方法在关键帧检测准确率、语义覆盖度、压缩比及实时性等方面表现优异. 消融实验进一步验证了 IMU 与语音模态在提升关键语义捕获能力方面的有效作用.

需要指出的是,本文方法主要适用于具有明显位姿变化和语音交互特征的移动终端场景. 对于固定视角、弱语音或低交互场景,其性能优势仍有进一步验证和提升的空间. 未来将围绕模型轻量化、跨场景泛化能力及更复杂环境下的鲁棒性开展进一步研究.

[参考文献]

[1] Sahu A, Chowdhury A S. Summarizing egocentric videos using deep features and optimal clustering[J]. Neurocomputing, 2020, 398: 209-221.

[2] Sahu A, Chowdhury A S. Egocentric video co-summarization using transfer learning and refined random walk on a constrained graph[J]. Pattern Recognition, 2023, 134: 109128.

[3] Hong F, Guzov V, Kim H J, et al. EgoLM: multi-modal language model of egocentric motions[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2025: 5344-5354.

[4] Moussaoui H, Akkad N, Benslimane M. A review of video summarization[C]//Proceedings of the 3rd International Conference on Digital Technologies and Applications. Fez, Morocco: Springer, 2023: 516-525.

[5] Zhao B, Gong M, Li X. Hierarchical multimodal transformer to summarize videos[J]. Neurocomputing, 2022, 468: 360-369.

[6] Patel A, Chitalia V, Yang Y. Advancing egocentric video question answering with multimodal large language models[J]. arXiv Preprint, arXiv: 2504.04550, 2025.

[7] Xie J, Chen X, Zhao S, Lu S P. Video summarization via knowledge-aware multimodal deep networks[J]. Knowledge-based Systems, 2024, 293: 111670.

[8] Tang X, Qiu J, Xie L, et al. Adaptive keyframe sampling for long video understanding[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2025: 29118-29128.

[9] Lee M J, Gong D, Cho M. Video summarization with large language models[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2025: 18981-18991.

[10] Meena P, Kumar H, Yadav S K. A review on video summarization techniques[J]. Engineering Applications of Artificial Intelligence, 2023, 118: 105667.

[11] Tan K, Zhou Y, Xia Q, et al. Large model based sequential keyframe extraction for video summarization[J]. arXiv Preprint,

arXiv:2401.04962,2024.

[12] Cakmak M C, Agarwal N, Poudel D. TriPSS: a tri-modal keyframe extraction framework using perceptual, structural, and semantic representations[J].arXiv Preprint,arXiv:2506.05395,2025.

[13] Li W, Qi D, Zhang C, et al. Video summarization based on mutual information and entropy sliding window method[J]. Entropy,2020,22(11):1285.

[14] Ranganath V, Padmavathi S, Aarthi R. Optimal frame extraction for event detection in surveillance videos[J]. Procedia Computer Science,2025,256:1192-1199.

[15] Arslan S, Tanberk S. Key frame extraction with attention based deep neural networks[J]. arXiv Preprint, arXiv: 2306.13176,2023.

[16] Kudo K, Nagasawa H, Suzuki J, et al. A challenging multimodal video summary: simultaneously extracting and generating keyframe-caption pairs from video[J].arXiv Preprint,arXiv:2312.01575,2023.

[17] 韩乾乾,顾华宁,金溢,等. 基于 X3D 特征和语义融合的篮球运动员检测跟踪方法[J].南京师大学报(自然科学版),2025,48(6):101-110.

[18] 何俊,张彩庆,李小珍,等. 面向深度学习的多模态融合技术研究综述[J].计算机工程,2020,46(5):1-11.

[责任编辑:杜忆忱]