AI 在提升视联网内容审核效率中的应用案例研究

陈卓玲[1] 任寒秋[2]（通讯作者）
1.中国电信股份有限公司浙江分公司 浙江杭州 310000；
2.中国电信股份有限公司宁波分公司 浙江宁波 315000

【摘 要】视联网内容呈井喷式增长态势，传统人工审核模式在海量内容实时安全管控方面已显力不从心。本研究着重探究AI技术于内容审核领域的创新性运用，搭建涵盖多模态融合分析、智能决策引擎及风险分级处置的技术架构，深度剖析计算机视觉、自然语言处理、深度学习等技术如何实现对视频、图像、音频、文本的智能化解析。以腾讯视频、字节跳动等企业实际操作为例，充分展现AI在色情内容识别、暴恐信息检测、违禁物品过滤等场景中的显著技术优势，研究发现，AI主导的“机器初步筛查与人工复核相结合”模式彻底重塑内容审核体系，为视联网平台打造智能化安全治理新生态提供极具借鉴意义的技术方案。

【关键词】人工智能；视联网；内容审核；效率提升；多模态分析

A case study on the application of AI in improving the efficiency of visual Internet content audit

Chen Zhuoling[1] Ren Hanqiu[2]（Corresponding author）
1.China Telecom Corporation Limited Zhejiang Branch Zhejiang Hangzhou 310000；
2.China Telecom Corporation Limited Ningbo Branch Zhejiang Ningbo 315000

【Abstract】The content of the visual internet is experiencing explosive growth，and traditional manual review models are struggling to manage the massive amount of content in real-time and securely. This study focuses on the innovative application of AI technology in content moderation，building a technical framework that includes multimodal fusion analysis，intelligent decision engines，and risk classification and handling. It delves into how computer vision，natural language processing，and deep learning technologies achieve intelligent parsing of videos，images，audio，and text. Using real-world examples from companies like Tencent Video and ByteDance，it fully demonstrates the significant technical advantages of AI in scenarios such as pornography detection，violent and terrorist information screening，and prohibited item filtering. The research finds that the AI-led model combining "initial machine screening with human re-review" has completely reshaped the content moderation system，providing a highly valuable technical solution for creating an intelligent security governance ecosystem on visual internet platforms.

【Key words】artificial intelligence；visual Internet of Things；content review；efficiency improvement；multimodal analysis

引言

步入视联网时代，短视频、直播等用户生成内容每日新增规模突破亿级大关，内容安全审核遭遇数量激增与类型繁杂的双重难题。传统人工审核模式受制于高昂成本、延迟的处理时效以及语义理解的局限性，难以契合实时且精准的治理要求，具备多模态数据处理能力与模式识别特长的 AI 技术，已然成为突破审核效率困境的关键力量。时下基于深度学习的计算机视觉手段可完成违规画面的动态捕捉，自然语言处理技术能够解读复杂语义情境，知识图谱联合动态决策引擎构建起智能化审核策略，本文期望为视联网平台搭建“技术驱动、人机协作”的新型审核架构，提供理论依据与实践指导，推动内容安全治理向数字化转型迈进。

一、AI 内容审核技术架构与核心能力构建

（一）多模态内容解析技术体系

互联网内容审核的复杂程度，与数据形态的丰富多样密切相关。AI 技术搭建起多层级多模态融合分析架构，以此达成对视频、图像、音频、文本的透彻理解，腾讯视频开发的智能审核系统，在视频内容处理上颇具建树，运用时空特征联合建模技术，借助 3D 卷积神经网络提取连续视频帧的空间与时间序列特征，搭配 LSTM 网络捕捉动作的时序关联，能够精确辨识打架斗殴、危险驾驶等动态违规场景。针对体育赛事直播里裸露肌肉易造成误判的情况，该系统引入人体姿态关键点检测模型，像 OpenPose，通过剖析关节点空间分布与运动轨迹，将暴力场景识别准确率从传统单帧检测的 65%跃升至 98%。图像审核层面，某直播平台打造的对抗增强识别模型，依托生成对抗网络训练鉴别器与生成器的对抗机制，使模型得以识别经马赛克处理、滤镜变形的隐蔽违规图像。实际操作时，生成器模拟违规内容的遮挡变形方式，诸如高斯模糊、色彩空间转换，鉴别器则学习区分真

实违规图像与生成图像的细微差别，经过 10 万轮对抗训练，涉黄图像漏检率从 12%下降到 2.3%，大幅增强了对复杂场景的处理效能。

自然语言处理技术在音频与文本审核领域有着深度实践，字节跳动直播音频审核体系配备方言自适应识别组件，对于川渝方言、粤语等复杂语音场景，运用基于注意力机制的双向 LSTM 网络开展声学模型训练，融入上下文关联音子建模策略，使得方言识别精度达到 92%，此系统嵌入环境噪声消除算法，借助短时傅里叶变换分离人声与背景杂音，搭配深度降噪自编码器还原清晰语音信号，让演唱会、户外直播等喧闹场景下违规语音识别准确率维持在 96%。于文本审核方面，基于 BERT 的语义理解模型经掩码语言模型与下一句预测任务预训练，可识别“福利”“约吗”等词汇于不同语境中的语义差别。某社交平台搭建的文本审核系统，针对“原味内衣”“成人玩具”等边缘商品描述，利用领域自适应微调技术在电商领域语料库二次训练，将违规文本分类准确率从传统关键词匹配的 75%跃升至 94%，成功攻克语义模糊场景下的审核困境。

（二）智能决策引擎与风险分级机制

AI 审核系统的关键优势在于搭建灵活可变的智能决策架构。某长视频内容平台打造的违规内容知识图谱，主要运用七元建模方式，将实体、关系、属性、时间、空间、规则及案例有机整合，把《网络安全法》、《网络视听节目内容审核通则》等法规条文转化为可运算的知识节点，构建起涵盖 2000 余个违规标签的分类体系。就“暴力内容”而言，其下延伸出“肢体冲突”“武器展示”“血腥画面”等 12 项子标签，每个子标签均对应专属视觉特征模型，像采用 HOG 特征结合 SVM 分类器检测刀具，运用光流法搭配 LSTM 识别肢体冲突，同时设定相应审核阈值，如将血腥画面的像素占比临界值定为 5%。

动态决策引擎依循内容风险评估模型，执行三级审核策略，面对低风险内容，诸如风景视频、教育讲座之类，AI 系统自动抓取关键帧，提取语音文本进行快速比对，短短 1 秒内便能完成审核并予以放行。中风险内容，像含有纹身画面的美妆视频、疑似出现脏话的游戏直播，会触发二次校验，AI 系统启动多模型交叉验证，同时启用肤色检测与姿态识别模型等，并将核验结果推送至人工审核工作台，要求在 3 分钟内完成复核。而高风险内容，包括暴恐视频、政治敏感画面，一旦出现便直接激活紧急阻断流程，于内容上传后 0.5 秒内迅速识别并拦截，同时启动 IP 地址追踪与用户行为分析模块，达成对风险源头的有效管控，某短视频平台运用此机制后，高风险内容平均处理时长由人工审核的 28 分钟锐减至 1.2 分钟，整体审核处理能力大幅提升 800%。

二、AI 在典型审核场景的应用实践

（一）色情内容智能识别与过滤

色情内容审核是 AI 技术较早实现大规模应用的场景之一，其主要面临的挑战在于对隐蔽化、变种化违规内容的识别，快手所研发的多任务学习检测模型，借助共享卷积神经网络底层特征提取层，能够同时输出人体关键点坐标、肤色区域掩码以及违规区域定位框这三个任务的结果，以此达成对裸露肢体、不当体位的联合检测[1]。在具体的技术路线方面，先将 ResNet50 用作主干网络来提取图像特征，而后通过三个分支网络分别处理姿态检测（输出 18 个关节点坐标）、肤色分割（生成 YCbCr 色彩空间的肤色概率图）、目标定位（输出违规区域的边界框），最后运用非极大值抑制（NMS）对多任务检测结果进行融合。该模型在诸如夜景直播、舞台灯光等复杂光照条件下的召回率可达到 94%，相较于传统的单任务模型提高了 9 个百分点。

某社交直播平台色情暗示检测系统于语义层面，以“语音识别 + 意图分类 + 语境建模”三级流程运作。ASR 技术把直播语音转文本，FastText 模型做意图粗分，区分正常交流、色情暗示等类别，Transformer 模型构建对话上下文语义向量，识别“来私密空间”“刷火箭看看”这类隐含色情邀请话术。系统结合画面实时分析，像镜头聚焦敏感部位时长、人体姿态异常角度，形成“语音语义 + 视觉特征”双重审核机制。最终实现涉黄内容拦截量月均提升 42%，人工审核工作量降低 55%。

（二）暴恐与违禁品内容检测

暴恐内容危害大且传播隐蔽，AI 审核系统借构建专业领域特征库和迁移学习技术解决小样本问题。某头部云厂商的视频审核暴恐特征数据库，有 10 万 + 暴恐图像，如极端组织标志、自制爆炸装置等。还有 5 万 + 极端主义视频，包含煽动性演讲、血腥处决画面[2]。用迁移学习在 ImageNet 预训练模型微调，针对“ISIS 旗帜”“AK47 步枪”等特定目标，以 Faster R-CNN 算法做目标检测，结合 3D 双流网络分析视频动作时序特征，实现对持械攻击、爆炸过程等暴恐行为序列的识别。

在违禁品检测领域，某云厂商为电商直播定制的审核系统，用 YOLOv5 算法做多尺度目标检测，对于口红、手表等商品里可能混有的管制刀具、仿真枪支，靠自适应锚框生成技术优化目标定位精度，能达到 0.1 秒 / 帧的实时检测速度。系统引入商品类别约束机制，在“美妆”直播间自动屏蔽刀具检测警报，防止误判形似化妆刷的物品，在“户外用品”直播间提高枪支识别阈值，实现检测准确率和业务场景的匹配，将违禁品漏检率控制在 0.5% 以下。部分企业针对暗网传播的加密暴恐内容，正探索基于区块链的哈希值比对技术，通过构建暴恐内容指纹库，实现对文件级违规内容的快速溯源。

三、AI 审核系统实施效果与优化路径

（一）效率提升与成本优化分析

AI 技术使视联网内容审核从劳动密集型转变为技术密集型，大幅提升审核效能。某长视频平台分布式 AI 审核系统，运用微服务架构和 GPU 集群并行计算，借助 Kubernetes 实现 10 万路视频流负载均衡[3]。单视频初筛包含关键帧提取、多模态特征融合、违规模型比对，200 毫秒就能完成，比人工审核每分钟 18 条且夜间效率下降 30% 的速度快很多，效率提升 600 倍以上。成本结构也发生变化，平台用 AI 取代基础审核岗位，审核人力成本从月均 800 万元降到 240 万元，降幅达 70%。原本 500 人的团队，350 人转到内容质量评估等需主观判断的环节，仅 150 人负责风险复核和数据标注，转型后释放的人力用于优化审核规则库，平均每年新增 300 + 细分场景规则，用户申诉响应时效提高 40%。

多模态融合技术打造三维审核体系，明显提升审核质量，视频审核时，3D-CNN 搭配肤色检测算法，使暴力内容识别准确率从 85% 升至 99.1%；音频审核依靠方言自适应和环境降噪技术，在嘈杂场景中脏话识别准确率从 75% 提高到 96%。总体而言，AI 审核系统把漏检率降到 2.8%，误判率控制在 4.5%，某短视频平台因误判的投诉量减少 62%。时效性上，AI 系统形成毫秒级闭环，就某短视频平台来说，基于 Flink 流处理引擎，用户点击“发布”后，系统 0-100ms 完成关键帧提取和哈希值生成，100-200ms 完成多模态特征向量计算，200-300ms 完成风险判定和处置指令生成[4]。98% 的违规内容在 500ms 内被拦截，而传统人工审核平均延迟 3 分钟，突发公共事件时 AI 优势突出，比如某明星负面事件发酵，AI 系统 1 秒内拦截 20 万条含谣言视频，比人工提前 20 分钟控制风险，防止舆情恶化。

（二）人机协同机制与技术迭代策略

为解决 AI 审核语义理解不足和文化差异问题，头部企业多构建“技术打底、人工校准、数据反哺”的闭环优化体系，抖音审核平台设实时交互模块，人工审核员处理 AI 初筛结果时，能修正误判案例标签、给模糊案例加专家注释。这些标注数据经脱敏后，每日自动同步到模型训练平台，实现违规特征库动态更新，平均每周新增 50 + 细分标签，像针对二次元文化“擦边球”内容，如过度暴露的动漫服饰，审核团队持续标注形成专属特征集，让系统识别准确率从初期 70% 提升到 88%。

在技术迭代方面，运用联邦学习来处理跨平台数据孤岛状况，各企业在不共享原始数据的情况下，借加密参数交换来协同训练模型，某联盟内企业对于小众违规内容，比如地方戏曲里特定服饰禁忌的识别能力平均提高 20%。主动学习算法的运用优化了数据标注效率，系统依靠不确定性采样，即选择模型置信度低于 70% 的样本，以及多样性采样，涵盖不同光照、角度、文化背景的样本，自动筛选出高价值标注样本，让人工标注成本下降 35%，标注效率提升 40%[5]。针对深度伪造技术导致的人脸替换、语音合成等新型违规内容，部分企业已展开对抗样本检测研究，通过生成对抗网络训练防伪模型，提取视频中生物特征一致性特征，像眨眼频率、微表情规律等，实现对伪造内容的有效辨别，相关技术检测准确率已达 91%，为应对技术滥用风险提供了前瞻性方案。

在伦理方面，AI 审核系统部署关注用户隐私保护和审核透明度构建。腾讯视频搭建审核日志区块链存证系统，对 AI 审核每次决策，像触发的模型、匹配的规则、拦截的依据等都上链记录，用户能通过哈希值查询审核具体情况，有效解决“审核黑箱”难题。系统运用差分隐私技术对用户数据做模糊处理，在保障个人信息的基础上开展模型训练，平衡了安全治理和用户权益的关系。这种技术创新与伦理考量兼顾的实践，为 AI 审核系统合规化部署提供了可借鉴的实施范例。

结语

AI 技术应用改变了视联网内容审核技术模式，借助多模态融合分析、智能决策引擎和风险分级处置，打造出高效准确的审核体系，实际情况显示，AI 极大提高了审核效率与精准度，还通过人机协同机制形成不断优化的治理循环。针对深度伪造技术、暗网内容传播等新问题，日后要进一步强化跨模态特征融合、小样本学习、对抗样本检测等技术研发，促使 AI 审核系统从“规则驱动”向“认知智能”转变。本研究为视联网平台建立智能化审核体系提供了能复制的技术路线，也为监管部门制定 AI 审核技术标准提供实践依据，有助于营造安全良好的视联网内容环境。

参考文献

[1]罗天.AI 在短视频内容审核中的自动分类与识别技术[J].家庭影院技术，2024，（10）：74-77.
[2]沈述宜，郭巧敏.算法时代的创意内容审核及生产——以短视频平台“快手”为例[J].美术大观，2023，（03）：144-147.
[3]陈晶，隗静秋，叶娜.网络短视频内容审核：基本遵循、现实困境与优化路径[J].江西科技师范大学学报，2023，（01）：82-88.
[4]牛百齐，王秀芳.人工智能导论[M].机械工业出版社：202301.240.
[5]邵恒媛.数字劳动视域下内容把关人工作的异化——基于今日头条人机协同审核机制的研究[J].媒体融合新观察，2021，（01）：38-43.