Considerations for Building a Multimodal Content Safety Review System (Postprint)

Authors: Liu Fan, Wang Fengmei

Date: 2023-10-08T00:00:00+00:00

Abstract

Objective: The rapid development of the Internet, smart devices, and various emerging services has led to massive Internet information being interspersed with large quantities of violent, sensitive, vulgar, and other undesirable information. With increasingly stringent national regulation of content safety, this study investigates methods to achieve rapid and precise content safety review of massive Internet information. Methods: Mainly utilizing big data and artificial intelligence technologies to innovate digital content review and filtering methods. Results: Achieving the integration of new technologies with traditional editorial review mechanisms. Conclusion: Transforming and upgrading from labor-intensive and brain-intensive models to innovation-intensive and technology-intensive models is an effective pathway and inevitable development trend for addressing the challenges of cross-modal content safety review in the media industry.

Full Text

Preamble

ChinaXiv Cooperative Journal
Thoughts on Constructing a Multimodal Content Security Review System
Liu Fan, Wang Fengmei
(Taiji Computer Corporation Limited, Beijing 100102)

Abstract

[**Purpose**] The rapid development of the Internet, smart devices, and various emerging services has resulted in massive amounts of online information intermingled with large quantities of violent, sensitive, vulgar, and other junk content. As national content security regulations become increasingly stringent,

this paper investigates methods for achieving rapid and accurate content security review of massive volumes of Internet information. [**Method**] The research primarily employs big data and artificial intelligence technologies to revolutionize digital content review and filtering methods. [**Result**] The study achieves integration of new technologies with traditional editorial review mechanisms. [**Conclusion**] Transforming from labor-intensive and brainpower-intensive models to innovation-intensive and technology-intensive models represents an effective approach and inevitable development trend for solving the cross-modal content security review dilemma in the media industry.

Keywords: multimodal; neural networks; big data; artificial intelligence; deep learning
Classification Code: G642
Document Code: A
Article ID: 1671-0134(2023)04-149-05
DOI: 10.19483/j.cnki.11-4653/n.2023.04.031
Citation Format: Liu Fan, Wang Fengmei. Thoughts on Constructing a Multimodal Content Security Review System [J]. China Media Technology, 2023(04): 149-153.

In today' s era of rapid media digitalization, the online information publishing environment has become increasingly complex, with content varying greatly in quality. Simultaneously, the arrival of the self-media era has brought explosive growth in content volume and variety, fundamentally transforming content production and distribution formats. Traditional content review and supervision methods consume substantially more resources while struggling to improve work efficiency. With the rapid development of the Internet, smart devices, and various emerging services, over 1 billion videos, images, and audio files are uploaded daily, alongside more than 500 million posts across social networks and media platforms—a trend that continues to accelerate. However, portions of this massive online information contain violent, sensitive, vulgar, and other junk content. As national content security supervision grows increasingly strict, traditional text proofreading and filtering systems can no longer meet the demands of mobile Internet content security review in this fast-paced era. Cross-modal content security review technology has become a serious challenge that the Internet industry must confront.

Traditional large websites have relied on manual review, where analysts examine content item by item—a process that guarantees neither efficiency nor accuracy. As artificial intelligence technologies mature, natural language processing, image recognition, and voiceprint identification have become applicable across most digital media domains. The rapid advancement of deep learning and natural language processing algorithms in AI has brought innovative solutions to these challenges, enabling not only precise identification of risky and sensitive information in content but also significantly reducing labor costs for content review. The introduction of AI technology will fundamentally transform traditional content review forms, enabling real-time review of Internet content and dramatically

improving both review efficiency and accuracy. Employing big data and AI technologies to revolutionize digital content review and filtering methods, and integrating high technology with traditional editorial review mechanisms, represents the transformation from labor-intensive and brainpower-intensive models to innovation-intensive and technology-intensive models—an effective approach and inevitable development trend for solving cross-modal content security review challenges in the media industry.

1. System Architecture Design

Multimodal content review constitutes the core capability of content security review systems. Leveraging industry-leading deep learning technology, natural language processing, optical character recognition (OCR), automatic speech recognition (ASR), and other technologies, the system provides intelligent risk identification services for multimedia content including images, videos, text, and audio. The service model offers flexible and diverse support for API/SDK, SaaS, private deployment, and other service methods while ensuring high availability and performance characteristics.

The multimodal content security review system is a complex system comprising numerous modules that can be organized into a typical hierarchical architecture by function, as shown in Figure 1 [Figure 1: see original paper].

Figure 1. Multimodal Content Security Review System Business Architecture

(1) Physical Infrastructure: A cluster of physical machines providing GPU (P4 recommended), CPU, memory, disk, network interface cards, and other physical resources.

(2) System Software: Software installed directly on physical machines. Virtualization and orchestration utilize open-source software Docker and K8S to virtualize physical machine clusters and provide orchestration APIs for upstream scheduling and usage. Upper-layer software is installed and deployed through K8S. Security authentication relies on plug-in dongle-based authentication services to ensure the security of multiple operators provided by the multimodal content security review system, particularly AI operators.

(3) Basic Software: Common foundational application software. This includes a log collection, retrieval, and viewing platform built on ELK; monitoring statistics, viewing, and alerting implemented through Prometheus+Grafana; a distributed file system built on Ceph supporting object storage, block device storage, and file system services; a high-availability MySQL cluster for databases; and middleware including high-availability Redis clusters and ZooKeeper.

(4) Operator Platform: Operators are the smallest computational units in the multimodal content security review system. Operator inputs can be a video, audio extracted from video, image sequences, outputs from other operators, or combinations thereof. The operator platform primarily provides automated

operations and maintenance for operators and operator task scheduling capabilities.

(5) General Platform: The general platform is not a single platform but a collection of relatively independent functional modules/subsystems that are not directly exposed to users but are relied upon by upper-layer business systems. Key subsystems include: - **Custom Face Recognition System:** Provides custom face database management, face feature computation (dependent on the operator platform), and face retrieval capabilities. - **Data Management Platform:** In the video AI all-in-one machine, input videos currently support only URL format. The data management platform provides data pulling, video metadata calculation, video transcoding, thumbnail extraction, audio extraction and VAD segmentation capabilities, while also supporting caching and reuse of intermediate video processing results. - **Evaluation Platform:** Using a batch of annotated data for testing and evaluation is a common method for judging all-in-one machine effectiveness. This process often requires manual initiation of predictions, analysis and comparison of prediction results, and statistical calculations—tedious operations that the evaluation platform streamlines. - **Statistics Platform:** The statistics platform provides service log query and business information statistics capabilities for the all-in-one machine. By collecting and storing service logs, it provides a unified log and statistical data query interface for users to troubleshoot issues and perceive business trend changes. - **Callback Service:** Provides the capability to send specific messages to specific addresses at specific times, supporting simple retry mechanisms and concurrency control.

(6) Business System: This system supports execution of different business scenarios for the multimodal content security review system (such as text analysis, image analysis, video analysis). Core functions include: - **Template Management:** Templates configure the types of operators users expect to apply to videos, with analysis tasks corresponding to different templates. - **Concurrency Control:** Based on cluster size, the business system needs to control the number of videos processed concurrently. - **Video Processing DAG Execution:** According to template configuration, the business system internally generates an operator execution roadmap (forming a DAG) for each video processing task, executing each operator sequentially (by calling the general platform and operator platform) to produce final outputs. - **Image Processing:** Supports image analysis and review 等业务 (business). - **Text Processing:** Supports text review 等业务 (business).

(7) Access Layer: Two access methods are provided: - **Console:** A visual operation interface for creating, modifying, and viewing templates, as well as initiating review tasks and viewing results. - **API/SDK:** Calls the business system' s Restful HTTP API for integration.

2. Review Capabilities and Typical Risk Scenario Design

2.1 Text Detection

Text detection leverages massive text feature libraries, rule libraries, keyword libraries, and NLP algorithms to filter and analyze text, helping content producers detect whether regulated text contains 违规信息 (non-compliant information). The system reviews multiple dimensions including pornography, terrorism, politics, advertising, prohibited content, insults, low-quality spam, negative comments, and ideological risk warnings, while supporting custom text blacklists. Typical risk scenarios that should be supported for recognition are described in Table 1 .

Table 1. Text Detection Typical Risk Scenario Description Table

Detection Type	Scenario Description
Spam Ads	Identifies text containing phone numbers, WeChat IDs, QQ numbers, URLs, check-ins, guided signatures, search prompts, and other information
Political Content	Identifies text involving negative political content, uncertain political content, specific individuals, character portrayals, events, or event dramatizations
Insults	Identifies text containing severe, moderate, or colloquial insults
Pornography	Identifies text including pornographic prohibitions, sexual knowledge, or suggestive content
Prohibited Content	Identifies text including part-time job offers, screen overlays, financial text messages, etc.
Custom Keywords	Identifies text matching custom keywords

2.2 Image Detection

Image detection applies active learning algorithms in artificial intelligence to quickly detect 违规内容 (non-compliant content) including pornography, political content, violence/terrorism, spam advertising, text-image violations, and image logos through deep learning models. Typical risk scenarios that should be supported for recognition are described in Table 2 .

Table 2. Image Detection Typical Risk Scenario Description Table

Detection Type	Scenario Description	Detection Results Classification
Intelligent Pornography Detection	Detects whether images contain pornographic or suggestive content	Normal, Pornographic, Suggestive
Violence/Terrorism/Politics Detection	Detects whether images contain violence/terrorism or political content	Normal, Violent/Terrorist/Political
Image QR Codes	Detects whether images contain QR codes or mini-program codes	Normal, Contains QR Code
Image Logos	Detects whether images contain logo information, such as station identifiers, trademarks	Normal, Contains Logo

2.3 Audio Detection

Audio content review helps content producers detect risks or 违规内容 (non-compliant content) in audio files or voice streams (such as live streams), including spam information, advertising, political content, violence/terrorism, insults, pornography, spam, prohibited content, and meaningless content. Typical risk scenarios that should be supported for recognition are described in Table 3 .

Table 3. Audio Detection Typical Risk Scenario Description Table

Detection Type	Scenario Description
Spam Ads	Detects audio containing phone numbers, WeChat IDs, QQ numbers, URLs, guided signatures, search prompts, check-ins, etc.
Political Content	Detects audio involving negative political content, uncertain political content, specific individuals, character portrayals, events, or event dramatizations
Insults	Detects audio containing severe, moderate, or colloquial insults
Pornography	Detects audio including pornographic prohibitions, sexual knowledge, suggestive content, or moaning sounds
Prohibited Content	Detects audio including part-time job offers, screen overlays, financial text messages, etc.

Detection Type	Scenario Description
Custom Keywords	Detects audio matching custom keywords

2.4 Video Detection

Video detection should distinguish between video files and live streams. Video files support default time-based frame extraction and user-defined frame extraction frequency after downloading via video URL parsing, followed by image detection and recognition on extracted frames. Live streams obtain video stream data through stream pulling, automatically converting video into images (frame extraction at set frequencies), then filtering and detecting the extracted images. The system should support filtering for both on-demand video and live streaming. The system should support both synchronous and asynchronous interface recognition access; asynchronous detection tasks do not return results immediately, requiring users to obtain detection results via Callback or polling methods. Typical risk scenarios that should be supported for recognition are described in Table 4 .

Table 4. Video Detection Typical Risk Scenario Description Table

Detection Type	Scenario Description	Detection Results Classification
Intelligent Pornography Detection	Detects whether video contains pornographic content	Normal, Pornographic
Violence/Terrorism/Politics Detection	Detects whether video contains violence/terrorism/political content	Normal, Violent/Terrorist/Political
Video Logo Detection	Detects whether video contains specific logos	Normal, Contains Logo
Video Text-Image Violations	Detects whether video contains advertising or 违规文字内容 (non-compliant text content)	Normal, Advertising or Text Violation
Video Voice Violations	Detects whether voice content in video contains 违规信息 (non-compliant information)	Normal, Contains Spam, Ads, Political, Violent/Terrorist, Insults, Pornographic, Spam, Prohibited, Custom (e.g., matches custom keywords)

3. Algorithm Recognition Capabilities and Key Technology Introduction

3.1 Pornography Recognition

Image pornography recognition utilizes knowledge from ultra-large-scale datasets to guide deep neural network training, producing multiple highly generalizable network models. Based on distillation learning principles, model complexity and parameter scale are significantly compressed to rapidly and accurately identify pornographic and vulgar images and videos, solving the problem of 违规内容识别 (non-compliant content identification).

3.1.1 Process Video detection first performs preprocessing to extract key image frames, converting the problem into image detection. Images extract features through pre-trained convolutional networks. Extracted features undergo binary classification through fully convolutional networks to determine whether they are pornographic or vulgar images. The process is illustrated in Figure 2 [Figure 2: see original paper].

Figure 2. Process Schematic

3.1.2 Principle Teaching machines to recognize pornographic images requires "training" them with thousands of image samples to extract and continuously memorize pornographic image features. Every point in an image includes brightness, hue, and saturation values. By setting value ranges for these three parameters, machines can identify "skin tones" and subsequently guess skin-exposed human body regions in images. The most obvious characteristic of pornographic images is the large proportion of human skin color in the frame. When machines identify skin-tone-like regions, they must further confirm the source of these regions—determining whether they are unclothed body parts or normal objects. Assuming two yellow regions represent legs or arms and another region represents the torso, if these regions' length and width values conform to human body proportions and their relative positions satisfy certain geometric relationships, the image is highly likely to be pornographic. If these regions' sizes and positions do not resemble a human body, the image can be excluded from pornographic suspicion. This geometric relationship calculation is illustrated in Figure 3 [Figure 3: see original paper].

Figure 3. Calculation of Skin Region Geometric Relationships

3.1.3 Classification Standards

- **Pornographic:** Images exposing sensitive body parts, containing explicit scenes, depicting sexual acts and pornographic scenarios
- **Suggestive:** Revealing clothing without exposing sensitive parts
- **Normal:** Non-pornographic, non-suggestive images

3.2 Violence/Terrorism Recognition

Image violence/terrorism recognition uses massive violence/terrorism image and video data sources, relying on distributed deep learning platforms to accurately classify images and videos. It specifically supports categories including bloody scenes, explosions, beheadings, parades and assemblies, fights and brawls, police-civilian conflicts, terrorism, war and military, guns and knives, sensitive clothing, sensitive text, and various flags.

3.2.1 Process Video first undergoes preprocessing to extract short video segments and key image frames. Short video features are extracted through "convolutional neural networks" and "recurrent neural networks," while video frame features are extracted through convolutional networks. Video features and image features are then fused and classified through fully convolutional networks and softmax classification functions to determine video categories. The process is illustrated in Figure 4 [Figure 4: see original paper].

Figure 4. Process Schematic

3.2.2 Principle Teaching machines to recognize violence/terrorism images similarly requires training with thousands of image samples to extract and memorize violence/terrorism image features. Each category of violence/terrorism images has obvious characteristic markers, such as the outlines of guns, daggers, and knives; flag pattern outlines; explosion scenes; color differences in bloody scenes, etc. Through continuous training, machines memorize these violence/terrorism features, enabling rapid feature value comparison when encountering new images to identify violent/terrorist content.

3.2.3 Classification Standards

- **Normal:** Images without violence/terrorism characteristics
- **Weapons/Weapon Holders:** Images showing guns, controlled knives, or their holders
- **Specific Individuals:** Images showing known terrorist leaders or politically sensitive figures
- **Special Symbols:** Special text appearing in images (including books), logos of violent/terrorist criminal organizations, some criminal television station logos, some religious symbols
- **Flags of Violent/Terrorist Organizations**
- **Specially Dressed Individuals:** Images featuring people wearing camouflage, military uniforms (including police, special police, and armed police), or special clothing
- **National Symbols:** Images containing a country' s national flag or emblem
- **Bloody Scenes:** Images showing bleeding, surgery, car accident blood, etc.
- **Riot Scenes:** Images showing parades, brawls, burning, etc.

- **War Scenes:** Images showing large-scale weapons (such as tanks, fighter jets), explosions, or groups of soldiers

3.3 Political Figure Recognition

Political figure recognition is based on massive face databases and professional reviewer standards, utilizing distributed deep learning platforms to identify 违规信息 (non-compliant information) involving normal, cartoon, or negative political figures, reducing 违规风险 (non-compliance risk). Coverage includes domestic and foreign heads of state, vice-national-level leaders and above, fallen officials, anti-China forces, and disgraced entertainers.

3.4 Text-Image Spam Advertising Recognition

Using deep learning algorithms combined with OCR technology and NLP natural language processing technology, the system identifies images, text, and watermarks in pictures to accurately detect spam content including QR codes, spam advertisements, pornography, political content, and insults.

Spam Advertisements: Images containing large amounts of solicitation, advertising, pornographic, or insulting text information.
QR Code Advertisements: Images containing printed QR codes or mini-program codes.

As media convergence advances deeply, with accelerated 5G deployment and the superposition of multiple emerging technologies such as big data, cloud computing, Internet of Things, blockchain, and artificial intelligence, mainstream central, provincial, and municipal/county-level media are using advanced technology as a core driving force to lead and drive integrated development, focusing on transforming toward smart media convergence. This has reshaped the ecological landscape of the media industry. How to ensure content security while intelligently and efficiently producing media convergence works has become a topic of common discussion among domestic and international media organizations, research institutions, and technology vendors. Currently, domestic high-quality artificial intelligence technology vendors, with years of product and technology accumulation, are intensifying research to explore how to better provide more reliable multimodal content security review products and services for media organizations.

References

[1] Wang Yahui, Wang Jing. The Transformation and Enlightenment of Artificial Intelligence on the Publishing Format of Scientific and Technology Journals [J]. China Media Technology, 2023(1): 52-55.
[2] Guo Yuhui. Research on False News Verification and Rating Mechanisms —Taking NewsGuard as an Example [J]. China Media Technology, 2022, (12): 29-32.

[3] Qiang Yanli. Research on the Impact of New Technologies on Media Formats and Media Digital Transformation [J]. China Media Technology, 2022(2): 103-105.
[4] Wang Zhengfang, Zhao Lei. Reconstructing the Communication Value Chain to Achieve All-Media Integrated Development—Thoughts on the Operation and Development of Radio and Television Stations [J]. Communication Power Research, 2018(29).
[5] Yu Guoming, Liu Yang. Media Production Innovation Based on Big Data in the Era of Media Convergence [J]. Media Observation, 2015(9).
[6] Wang Hui. The Impact and Changes of Artificial Intelligence Technology on the Broadcasting and Hosting Industry [J]. Media Forum, 2019(9): 120.
[7] Zhang Hao, Wu Jianxin. A Survey of Unsupervised Image Retrieval Research Based on Deep Features [J]. Journal of Computer Research and Development, 2018(9): 1829-
[8] Invention Patent "A Video Content Review System and Method" [P]. Patent Number: CN200610167182.7.
[9] Song Qing, Qi Chenglin, Zhang Pengzhou. Thoughts on the Application of Knowledge Graph Technology in the News Field [J]. China Media Technology, 2016(5): 19-21.

Author Biographies:

Liu Fan (1979-), male, from Suzhou, Anhui, senior professional title, Assistant President of Taiji Computer Corporation Limited, research direction: media convergence.
Wang Fengmei (1988-), female, from Shandong, intermediate professional title, Taiji Computer Corporation Limited, research direction: media convergence, media big data applications.

(Responsible Editor: Zhao Guoxu)

Note: Figure translations are in progress. See original paper for figures.