HiLL PUBLISHING

Advances in Computer and Communication, 2024, 5(3), 200-204
https://www.hillpublisher.com/journals/acc/
ISSN Online: 2767-2875

Research on the Construction and Practice of Intelligent Identification of Webcast Behavior

Feng Chen

Zhejiang Business Technology Institute, Ningbo 315012, Zhejiang, China.

How to cite this paper: Feng Chen. (2024) Research on the Construction and Practice of Intelligent Identification of Webcast Behavior. *Advances in Computer and Communication*, 5(3), 200-204.
DOI: 10.26855/acc.2024.07.008

Received: June 7, 2024
Accepted: July 4, 2024
Published: August 2, 2024

***Corresponding author:** Feng Chen, Zhejiang Business Technology Institute, Ningbo 315012, Zhejiang, China.

Abstract

This research focuses on the intelligent identification of live behavior and aims to build an efficient and accurate live behavior identification system. The paper first reviews the existing webcast behavior recognition technology and analyzes its advantages and disadvantages. Based on this, an intelligent recognition model of live behavior based on deep learning and multimodal fusion is proposed. The model uses multi-source data such as video, audio, and text to extract features through the deep neural network, and uses an attention mechanism to realize the effective fusion of multi-modal information. To verify the effectiveness of the model, the research team constructed a large-scale data set of webcast behavior, covering a variety of live broadcast scenarios and behavior types. Experimental results show that the proposed model outperforms existing methods in both recognition accuracy and real-time performance. Finally, the model is applied to the actual network broadcast platform, and its practical value in content review and user behavior analysis is discussed. This study provides new ideas and methods for the intelligent identification of network broadcast behavior, which is of great significance for improving the management level and user experience of live broadcast platforms.

Keywords

Network broadcast; behavior recognition; deep learning; multi-modal fusion; intelligent audit

In recent years, with the rapid development of Internet technology and the popularity of mobile devices, webcasts have become a widely popular way of information dissemination and social entertainment. However, the diversity and real-time nature of live broadcast content also bring great challenges to platform management and content audit. The traditional manual audit method has made it difficult to meet the regulatory needs of massive live broadcast content, so it is particularly important to develop efficient and accurate automatic live broadcast behavior recognition technology. Different from traditional video, network broadcast has the characteristics of strong real-time, complex scenes, and rich information mode, which brings new problems for behavior recognition. At the same time, the identification results of live broadcasting behavior are directly related to the content security and user experience of the platform, so there are high requirements for the accuracy and real-time performance of the identification. By comprehensively using deep learning, computer vision, and natural language processing technologies, a multimodal integrated live broadcast behavior recognition model is constructed to improve the accuracy and robustness of recognition. At the same time, the research will also discuss the application of this technology in the actual live broadcasting platform, to provide technical support for improving the management level of live broadcasting content and user experience.

1. The concept of intelligent recognition of webcast behavior

Intelligent identification of network broadcast behavior refers to the process of automatic analysis and identification of anchors and viewers in the use of artificial intelligence and machine learning technology. Through the analysis of the content, language, image, sound and other data of the live broadcast, the nature, characteristics and risk degree of the live broadcast behavior can be judged, so as to realize the management and monitoring of the network live broadcast platform. The collected data is processed and analyzed to extract the key features, such as the emotional color of the sound, the content characteristics of the image, the emotional tendency of the text, etc. Machine learning and deep learning algorithms are used to build webcast behavior classification and identification models [1]. The network live broadcast behavior is divided into different categories, such as normal live broadcast, illegal live broadcast, infringement live broadcast, etc. The classification model can classify and label the live broadcast behavior according to the extracted features. The risk of the divided live broadcast behaviors is assessed to determine the degree of potential threat to users and the platform. Risk analysis and prediction can be performed based on the characteristics of the behavior and historical data. Based on the identification results, the network live broadcast platform can formulate corresponding management strategies and measures, timely find and deal with violations, protect the rights and interests of users, and maintain the order of the platform [2].

2. The main content of network broadcast behavior identification

2.1 Analysis of the behavioral characteristics of the network broadcast

Webcast behavior has unique characteristics which have important influence on the choice of methods and techniques for behavior recognition. First of all, the network broadcast has a strong real-time nature, and the live broadcast content is generated and transmitted in real time, which requires the identification system to quickly process and analyze the data. Secondly, the live broadcasting scenes are diversified, including games, music, education, e-commerce and other fields, each of which has its own specific behavior patterns and language characteristics [3]. Thirdly, live broadcasting behavior usually involves a variety of modal information, including video image, audio voice and text barrage, etc. These information complement each other and together constitute the complete performance of live broadcasting behavior. In addition, live broadcasting behavior is also characterized by strong interaction, with frequent interaction between anchors and viewers, and these interactions are also important clues for behavior recognition. From the perspective of time dimension, live broadcast behavior can be divided into instantaneous behavior and continuous behavior. Instantaneous behaviors such as bowing, and waving, and continuous behaviors such as dancing, explaining, etc. From the perspective of spatial dimension, we can pay attention to the facial expressions, body movements, and the environment of the anchor. From the semantic level, it is necessary to understand the language content, emotional expression and audience feedback of the anchor. These multi-dimensional features provide a rich source of information for behavioral recognition but also increase the complexity of recognition.

2.2 Computer vision and image recognition technology

Computer vision and image recognition technology play a core role in the behavior recognition of webcasts, mainly responsible for processing and analyzing the visual information in video streams. The technology in this field has developed rapidly, from the early feature engineering based image to modern deep learning methods, which continuously improves the accuracy and efficiency of image recognition. In terms of feature extraction, traditional methods such as SIFT and HOG still play a role in some scenarios, while deep learning-based methods such as convolutional neural network (CNN) have become the mainstream [4]. Various CNN architectures such as VGGNet, ResNet, and Inception have achieved remarkable results in tasks such as image classification, target detection, and semantic segmentation. For the processing of video sequences, recurrent neural network (RNN), long and short-time memory network (LSTM), and 3D convolutional network are widely used to capture temporal information. In recent years, attention mechanism and Transformer architecture have also made breakthroughs in the field of computer vision, such as the Vision Transformer (ViT) model showing powerful performance in image classification tasks. Moreover, target detection and tracking techniques such as YOLO, SSD, and Faster R-CNN provide powerful tools for identifying specific objects and actions in live broadcasting. In the recognition of webcast behavior, these technologies are used to analyze visual information such as facial expressions, body movements, and the environment of anchors. For example, the emotional state of anchors can be analyzed by facial expression recognition technology, and specific

movements and gestures can be identified by gesture estimation technology [5]. At the same time, scene understanding technology can help identify the scene types of live broadcasting and provide context information for behavior recognition.

2.3 Natural language processing technology

Natural language processing (NLP) technology is mainly used to analyze and understand the voice content and text information in the process of live broadcast in the behavior recognition of network live broadcast, including the voice of anchors and the audience's barrage comments, etc. In recent years, NLP technology has made remarkable progress, providing a powerful tool for live broadcast behavior recognition. In terms of speech recognition, end-to-end deep learning models such as DeepSpeech and Wav2Vec greatly improve the accuracy of speech-to-text. These technologies can convert the real-time speech of anchors into text, providing the basis for subsequent semantic analysis. For converted text or direct bullet screen text, various advanced NLP models such as BERT, GPT, XLNet, etc. can be used for deep semantic understanding. Through pre-training and fine-tuning, these models can effectively capture the contextual information and deep semantics of the text. Emotion analysis is another important application of NLP in live broadcast behavior recognition. By analyzing the language expression of anchors and audience comments, the emotional atmosphere and interactive quality of live broadcasts can be evaluated [5]. Named entity recognition (NER) technology can extract key information from the text, such as human names, place names, product names, etc., which is helpful in understanding the theme and content of the live broadcast. In the context of network broadcast, the main challenges facing NLP technology include high real-time requirements, multiple colloquial expressions, frequency of new words, and multilingual mixing. Therefore, the research focuses include improving the processing speed of the model, enhancing the understanding ability of spoken language and network language, and improving the cross-language ability of the model.

2.4 Deep learning and neural network

Deep learning and neural network technologies play a central role in webcast behavior recognition and provide powerful tools for processing complex multimodal data. These techniques are able to automatically learn feature representations, greatly reducing the requirements of manual feature engineering, while improving the performance and generalization capability of the model. In the deep learning framework, the convolutional neural network (CNN) is mainly used to process image and video data. Various advanced CNN architectures such as ResNet, DenseNet, and EfficientNet can not only effectively extract visual features, but also process large-scale data sets. For the processing of temporal data, recurrent neural network (RNN) and their variants such as Long-short Memory Network (LSTM) and gated cycle unit (GRU) are widely used to analyze speech and text sequences. In recent years, the attention mechanism and the Transformer architecture have made breakthroughs in various fields. In live broadcast behavior recognition, these techniques can be used to capture long-distance dependencies between different modalities, such as BERT for text comprehension, and Vision Transformer for image analysis. Deep reinforcement learning techniques are also beginning to play a role in behavior recognition, especially when dealing with complex interactive behaviors. For example, reinforcement learning can be used to optimize multimodal fusion strategies or dynamically adjust feature extraction concerns. Another important research direction is the lightweight and acceleration of the model to meet the real-time processing requirements in live broadcast scenarios [6]. Techniques such as model compression, knowledge distillation, and quantization are used to reduce the computational complexity of the model while maintaining performance.

2.5 Multimodal information fusion method

Multi-modal information fusion is a key technology in the identification of webcast behavior, which aims to effectively integrate information from different perceptual channels (such as visual, auditory, and text), in order to obtain a more comprehensive and accurate understanding of live broadcast behavior. The main challenges of multimodal fusion lie in how to handle the heterogeneity of data, timing alignment of different modes, and how to fully utilize the complementarity and redundancy among different modes. From the level of fusion, multimodal fusion can be divided into early fusion, late fusion, and mixed fusion. The early fusion is performed directly out at the feature level, such as the simple splicing of feature vectors of different modes; the late fusion is conducted at the decision level, and each mode is integrated; the hybrid fusion is performed at the middle level, allowing information interaction at different levels. In specific fusion methods, attention mechanisms are widely used in multimodal fusion [7]. For

example, self-attention mechanisms can be used to capture long-distance dependencies within a single modality, whereas cross-attention mechanisms are used to learn associations between different modalities. Transformer Architecture and its variants, such as MMBT (Multimodal Bitransformer), ViLBERT, etc., realize efficient multimodal fusion through the self-attention mechanism. Another important kind of fusion method is that based on graph neural network (GNN) method. By representing the information of different modalities as nodes of the graph, GNN can effectively model complex relationships between modes. For example, a dynamic fusion graph network (Dynamic Fusion Graph Neural Networks) can adaptively learn the fusion weights between different modes. In recent years, multimodal pre-training models such as CLIP and DALL-E have shown strong cross-modal understanding ability, providing new ideas for live broadcast behavior recognition.

3. Design of the intelligent identification model of network broadcast behavior

3.1 Video feature extraction module

The video feature extraction module is a key component of the intelligent recognition system of network broadcast behavior, which is responsible for extracting meaningful visual features from the video streams. These features include, but are not limited to, the anchors' facial expressions, body movements, scene information, etc. In the preprocessing phase, the video streams are first decoded and split into consecutive series of frames. To reduce the computational burden, frame sampling is often performed, such as sampling a fixed number of frames per second. Samsampled frames are standardized, including sizing, color space conversion (e.g., RGB to BGR), and pixel value normalization. For static feature extraction, a convolutional neural network (CNN) is the main tool. Pre-trained CNN models such as ResNet, Inception, or EfficientNet can be used to extract high-level visual features for each frame. These models are typically pre-trained on large-scale image datasets and then adapted to live scenes by fine-tuning. To capture temporal information, a 3D convolutional network (C3D) or its variants such as I3D (Inflated 3D ConvNet) can be used. These models are able to learn spatiotemporal features directly from continuous video segments. Another approach is to use a two-stream network, one stream processes the RGB images and the other stream processes the optical stream information and then fuses the features of the two streams. For specific tasks, such as facial expression recognition or pose estimation, specialized models can be introduced. For example, the facial keypoint detection model can be used to extract facial features, while the human pose estimation model such as OpenPose can be used to capture the action information of anchors.

3.2 Audio feature extraction module

The audio feature extraction module plays an important role in the intelligent recognition system of network broadcast behavior and is mainly responsible for extracting valuable acoustic features from the audio stream of the live broadcast. These characteristics can reflect the voice content, emotional state, background environment, and other information [8]. In the preprocessing stage, the audio signal is first noise-reduced and framed. Commonly used noise reduction techniques include spectrum subtraction and Wiener filtering. Framing is usually performed with a fixed length window (e. g. 25ms) and a step size (e. g. 10ms). Traditional audio feature extraction methods are still widely used, such as Mayer frequency inversion coefficient (MFCC), linear prediction coefficient (LPC), over zero rate (ZCR), etc. These features can effectively capture the spectral characteristics and time domain characteristics of audio. For music live streaming, musicology features such as color features (Chroma Features) can also be extracted. The latest research trend is to use self-supervised learning methods to learn audio representations. For example, models such as wav2vec and HuBERT learn powerful audio feature representation through pre-training of large-scale unannotated data, which can be applied to live behavior recognition tasks through fine-tuning. Real-time performance is an important consideration in live broadcast scenes. Therefore, we need to weigh the complexity of feature extraction against the response time of the system. Consider using flow processing techniques such as online ASR systems to achieve near real-time feature extraction. The output of the audio feature extraction module is usually a sequence of feature vectors, each vector corresponding to an audio frame or segment. These features will be sent into the multimodal fusion module for integration with the other modal features. Design of effective audio feature extraction modules is crucial for accurate understanding of live content and behavior.

3.3 Text feature extraction module

The text feature extraction module mainly processes two types of text data in the intelligent behavior recognition

system of network broadcast: the bullet screen comments during the live broadcast and the anchor voice text obtained through speech recognition conversion. The main goal of this module is to extract meaningful semantic features from these texts to support subsequent behavioral recognition tasks. In the pre-processing stage, the text is cleaned first, including the removal of special characters and emojis. For Chinese text, word segmentation processing is also required. Given the extensive use of network terms and new words in live scenes, customized specialized word splits and dictionaries may be required [9]. Traditional text feature extraction methods such as TF-IDF and pouch models are still used in some scenarios, especially when computational resources are limited or when rapid processing is needed. These methods can quickly capture the keyword distribution of the text. Word embedding techniques are the foundation of modern NLP, and methods such as Word2Vec, GloVe can map words to a low-dimensional dense vector space. For live scenes, domain-specific pre-trained word vectors can be considered to better capture the characteristics of live language.

4. Conclusion

Intelligent identification of webcast behavior is a complex and challenging research field, involving many aspects such as multimodal information processing, feature extraction, fusion, and classification. This paper discusses in detail the overall architecture of the system, including video, audio and text feature extraction modules, as well as multimodal feature fusion and behavioral classification identification modules. With the continuous development of deep learning and artificial intelligence technologies, especially the advances in multimodal learning and cross-modal understanding, the performance of webcast behavior recognition systems is expected to be further improved. With the progress of technology and the expansion of application scenarios, the intelligent recognition system of network broadcast behavior is expected to play an important role in content review, personalized recommendation, interactive experience optimization, and other fields, providing strong support for the healthy development of the live broadcast industry.

Funding

This paper is supported by the second batch of Ningbo philosophy and social science planning project in 2023 (annual application project): online live streaming * risk prevention path research—based on the perspective of Ningbo live streaming industry development (G2023-2-66), Moderator: Feng Chen; Ningbo 2023 year of the second batch of social public research project: based on ontology and NLP * identification key technology research and application (2023S169), the host: Feng Chen.

References

[1] Li Ming, Wang Qiang, and Zhang Wei. Research progress in Video Behavior Recognition based on Deep Learning. Journal of Computer Science, 2022, 45 (3): 521-544.

[2] Chen Yan, Liu Zhi, and Zheng Yu. Review of multimodal emotion recognition techniques. Journal of Automation, 2023, 49 (2): 201-220.

[3] Zhao Wenbo, Huang Ming, Chen Xiaohong. Real-time voice and emotion recognition method in webcast. Computer Research and Development, 2021, 58 (4): 812-825.

[4] Wang Li, Li Hua, and Zhang Ming. Research on the multimodal feature fusion method based on the attention mechanism. Pattern Recognition and Artificial Intelligence, 2022, 35 (1): 67-79.

[5] Liu Qian, Sun Hong, Zhou Zhihua. Summary of text classification and emotion analysis in live broadcast scenes. Journal of Software, 2023, 34 (3): 1-20.

[6] Zhang S, Li S, and Wang W. Research on the application of transfer learning in webcast Behavior Identification. Chinese Science: Information Science, 2021, 51 (8): 1289-1305.

[7] Daniel W, Edison C, and Andy L. Multimodal feature fusion method based on graph neural network. Journal of Computer-Aided Design and Graphics, 2022, 34 (5): 789-801.

[8] Lin ZL, Fan BB, Yang M. Research on real-time human pose estimation technology under live broadcast scene. Chinese Journal of Graphic Pictures, 2023, 28 (2): 345-358.

[9] J Chou, JJ Lin, and LH Wang. Multi-task learning method in webcast content review. Journal of Communications, 2022, 43 (7): 112-126.